

OXFORD



EDUCATIONAL MEASUREMENT

FIFTH EDITION

EDITED BY

LINDA L. COOK & MARY J. PITONIAK

Educational Measurement

Educational Measurement

FIFTH EDITION

Edited by

**Linda L. Cook
and Mary J. Pitoniak**

Sponsored by National Council
on Measurement in Education



OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide. Oxford is a registered trademark of Oxford University Press in the UK and in certain other countries.

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

© National Council on Measurement in Education 2025

This is an open-access publication, available online and distributed under the terms of a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International license (CC BY-NC-ND 4.0), a copy of which is available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>. Subject to this license, all rights are reserved.

Fifth Edition published in 2025



Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

CIP data is on file at the Library of Congress

ISBN 9780197654965

DOI: 10.1093/oso/9780197654965.001.0001

Printed by Integrated Books International, United States of America

We were both fortunate to have Ron Hambleton as our advisor at the University of Massachusetts Amherst. We would like to acknowledge Ron as a teacher, as a leader in the field, and as a friend. His determination to advance the science of educational measurement through his teaching and his research was extraordinary. We shall forever be thankful for what he gave to us in both our professional and our personal lives.

Linda would also like to dedicate the book to her dear husband, Daniel Eignor. Dan was both a friend and a colleague to her. She is grateful for his support and friendship in the years they were able to have together.

Mary would also like to thank her parents, Thomas and Catherine Pitoniak, for their support of her education and their belief in her capabilities. She also thanks her canine companions throughout the years for their unconditional support of psychometrics.

Linda L. Cook
Mary J. Pitoniak
Coeditors

Contents

List of Tables	XI
List of Figures	XV
Foreword	XIX
Acknowledgments	XXI
Contributor List	XXIII

PART I: SETTING THE CONTEXT

CHAPTER 1	Factors Influencing Education and Educational Measurement	3
	<i>Linda L. Cook and Mary J. Pitoniak</i>	
CHAPTER 2	The History of Educational Measurement	33
	<i>Brian E. Clauser, Jerome C. Clauser, and Amanda L. Clauser</i>	
CHAPTER 3	The Sociocultural Context of Educational Assessment	115
	<i>Kadriye Ercikan and Guillermo Solano-Flores</i>	

PART II: FOUNDATIONAL CONCEPTS

CHAPTER 4	Validity and Validation	191
	<i>Suzanne Lane and Scott F. Marion</i>	
CHAPTER 5	Reliability in Educational Measurement	277
	<i>Won-Chan Lee and Deborah J. Harris</i>	
CHAPTER 6	Fairness in Educational Measurement: Theory and Concepts	383
	<i>Rebecca Zwick</i>	

PART III: DEVELOPING, ADMINISTERING, AND SCORING ASSESSMENTS

CHAPTER 7	Designing and Developing Educational Assessments	441
	<i>Kristen Huff, Paul Nichols, and M. Christina Schneider</i>	
CHAPTER 8	Test Administration and Scoring	513
	<i>Mark D. Shermis, Allan S. Cohen, and Jordan M. Wheeler</i>	
CHAPTER 9	Technology-Based Assessment: Validity, Modeling, and Analysis Issues	581
	<i>Randy E. Bennett, Michelle LaMar, and John Mazzeo</i>	
CHAPTER 10	Modeling Data From Educational Assessments	655
	<i>Li Cai, Mark Hansen, Minjeong Jeon, and Michael C. Edwards</i>	
CHAPTER 11	Scaling, Equating, and Linking	735
	<i>Tim Moses</i>	
CHAPTER 12	Standard Setting: A Cognitive and Social Model	821
	<i>Steve Ferrara, Susan Davis-Becker, Priya Kannan, and Katherine Reynolds</i>	
CHAPTER 13	Reporting Scores and Other Results	895
	<i>April L. Zenisky, Francis O'Donnell, and Ronald K. Hambleton</i>	
CHAPTER 14	Realizing Fairness Through Accessibility for All Test Takers and for Specific Groups	959
	<i>Michael C. Rodriguez and Martha L. Thurlow</i>	

PART IV: SPECIFIC APPLICATIONS OF ASSESSMENT

CHAPTER 15	Assessment to Inform Teaching and Learning	1017
	<i>Susan M. Brookhart and Charlie DePascale</i>	
CHAPTER 16	Test-Based Accountability in K–12 Education	1113
	<i>Andrew D. Ho and Morgan S. Polikoff</i>	
CHAPTER 17	Assessment for Admission, Placement, and Outcomes in Higher Education	1189
	<i>Wayne J. Camara, Ou Lydia Liu, and Krista Mattern</i>	
CHAPTER 18	Assessment for Licensing and Certification	1265
	<i>Melissa J. Margolis, Rebecca S. Lipner, and Chad W. Buckendahl</i>	

CHAPTER 19	Assessment of Social-Emotional, Soft, Character, Behavioral, and Intrapersonal and Interpersonal Skills	1339
	<i>Patrick C. Kyllonen and Jiyun Zu</i>	
CHAPTER 20	Large-Scale Assessments in International Contexts	1423
	<i>Henry I. Braun and Irwin Kirsch</i>	
 PART V: MEASUREMENT AS A SCIENCE		
CHAPTER 21	On the Nature of Measurement	1513
	<i>Derek C. Briggs, Andrew Maul, and Joshua A. McGrane</i>	
	 Index	 1579

List of Tables

4.1	Messick's Facets of Validity	205
4.2	Interpretation and Use Argument and Validity Argument Components for Culturally Responsive Assessments	247
4.3	Interim Assessment Uses	250
5.1	Variance and Covariance Components for the Extended Multivariate Design Example	308
5.2	Variance-Ratio Reliability Coefficients for the Maximum Likelihood Estimator	315
5.3	Estimators of Conditional Standard Errors of Measurement Based on a Single Test Administration	336
7.1	Various Characterizations of Cognition Related to Assessment	457
7.2	Example Range Performance-Level Descriptors	464
7.3	Claims and Evidence in Principled Assessment Design	487
7.4	Validation Framework and Criteria for Performance-Level Descriptors, Design Patterns, and Item Development	490
8.1	Comparison of English-Based Automated Essay Scoring (AES) Systems	535
8.2	Comparison of Non-English-Based Automated Essay Scoring (AES) Systems	559
10.1	An Overview of More Specialized Models in the General Framework	661
10.2	Descriptive Statistics of Sum Scores of Pretest and Posttest by Condition	701
10.3	Hierarchical Linear Model Results Based on Test Sum Scores	703
10.4	Classical Item Statistics	704
10.5	Unidimensional Item Response Theory Model Item Parameters	705
10.6	Unidimensional Item Response Theory Model: Marginal X^2 and Standardized Local Dependence X^2 Statistics	706
10.7	Model Fit Statistics: Unidimensional Item Response Theory Model Versus Bifactor Model	708
10.8	Model Fit Statistics: Single-Level Bifactor Model Versus Two-Level Bifactor Model	709
10.9	Q-Matrix Specified for Diagnostic Classification Models	710
10.10	Top Five Estimated Proportions of Attribute Profiles in Diagnostic Classification Models	711
10.11	Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Data Structure	712
10.12	Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Factor Pattern	713

10.13	Descriptive Statistics of Latent Pretest Score and Latent Change Score by Condition	715
10.14	Two-Level Item Response Theory Growth Model With Condition Indicator as a Predictor: Data Structure	715
10.15	Two-Level Item Response Theory Growth Model With Condition Indicator as a Predictor: Key Results	716
10.16	Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Data Structure	717
10.17	Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Key Results	717
10.18	Summary of Alt ELPA Grade 2–3 Item Characteristics	719
10.19	Model Fitting Summary and Key Results: Multidimensional Item Response Theory (IRT) Model Versus Cross-Classified IRT Model	720
10.20	Cross-Classified Item Response Theory Model: Data Structure (Example 2)	721
11.1	Summary of Data Collection Designs for Scaling, Equating, and Linking	738
11.2	Summary of Types of Equating	761
11.3	Summary of Nonequating Types of Linking	787
11.4	Summary of Recent Linking Examples	787
12.1	Selected Milestones in the History of Standard Setting	825
12.2	Taxonomy of Cognitive-Judgmental Tasks With Key Elements of Eight Standard-Setting Methods (and Two Variations), Including Paraphrased Statements of Cognitive-Judgmental Tasks	835
12.3	Framework to Guide Workshop Design and Gather Validation Evidence to Support Claims About the Interpretation of Cut Scores	857
12.4	Proposed Proactive Training Strategies to Reduce Deleterious Effects on Panelist Judgment and Decision-Making of Judgmental Heuristics and Cognitive Biases (or “Debiasing”)	862
13.1	Updated Evaluation Form for Reviewing Reports	932
15.1	Framework Relating Formative Assessment Strategies to the Formative Assessment Process and Its Agents	1021
15.2	Examples of Classroom Formative Assessment	1024
15.3	Construct Map Showing a Learning Progression for Student Understanding of Earth in the Solar System	1065
16.1	An Example of a Categorical Model From Delaware’s 2009–2010 School Year	1151
17.1	ACT and SAT Content, Testing Time, and Item Counts	1207
17.2	Percentage of U.S. Applications and Enrollment by Selectivity at 4-Year Institutions for 2017–2018	1209
17.3	Estimated Number and Percentage of U.S. High School Graduates Taking College Admissions Tests in 2017	1212

17.4	Graduate and Professional Admissions Tests	1216
17.5	Reliability of Undergraduate and Graduate Admissions Tests	1223
17.6	Validity of Undergraduate and Graduate Admissions Tests for Predicting First-Year Grade Point Average (GPA)	1225
17.7	Estimated Percentage of U.S. Postsecondary Institutions Using Selected Math and Reading Tests for Placement in Fall 2011	1235
19.1	Principal Component Analysis of 21st-Century Skills and Big Five Alignment	1349
20.1	School-Based International Tests of Educational Achievement: Scope and Timing	1428
20.2	Purposes of School-Based International Assessments	1473
21.1	The Politics of the National Assessment of Educational Progress (NAEP) Reading Framework Revision	1565

List of Figures

3.1	The Bidirectional Relationship Between Sociocultural Context and Assessment	117
3.2	Relationship Between Large-Scale Assessment and Society	128
3.3	The Assessment Triangle	152
3.4	Factors and Considerations Impacting the Cognition Model	153
3.5	Factors and Considerations Impacting the Observation Model	155
3.6	Factors and Considerations Impacting the Interpretation Model	159
3.7	Sociocultural Considerations in the Assessment Triangle	160
3.8	Conditional Inference About Student Knowledge and Abilities	166
4.1	Inferential Links in the Assessment Use Argument	232
7.1	Distinguishing Characteristics of Principled Assessment Design	446
7.2	Principled Assessment Design Essential Elements	453
7.3	National Assessment of Educational Progress Grade 4 Mathematics Policy-Level Definitions	461
7.4	Range Performance-Level Descriptor Examples	462
7.5	Remaining Steps in Assessment Design	465
7.6	Design Pattern Organization and Grain Size	466
7.7	Snapshot of Design Pattern With Manipulable Features of Complexity (Difficulty Drivers)	469
7.8	Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers English Language Arts/Literacy	470
7.9	Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers Math	471
7.10	Conventional Test Specifications With Separate Considerations of Components	475
7.11	Snapshot of Integrated National Assessment of Educational Progress Technology and Engineering Literacy Specifications	476
7.12	Hypothetical Comparison of Intended Performance Levels and Item Scale Placement	478
7.13	Generating Performance-Level Descriptors for the Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores Partnership	481
7.14	Sample Theory of Action for Interim Assessment	484
7.15	The Nested and Hierarchical Nature of Claims	488
7.16	Item Development Checklist Example	492

7.17	Example Item in English and Spanish to Illustrate Accessibility Features	496
8.1	An Example of a Hotspot Question	529
8.2	An Example of a Drag-and-Drop Question	530
8.3	A Screenshot of a Video Prompt	531
8.4	A Screenshot of Dissection Simulator Frog Edition	532
8.5	A Screenshot From the Serious-Based Game <i>80 Days</i>	533
8.6	Organization and Construct Coverage of e-rater Version 19.1, Released in 2019	537
8.7	Standard Process Approach for Automated Speech Scoring	539
8.8	An Example of a Document-by-Word Matrix	543
8.9	Illustration of Singular Value Decomposition on a Document-by-Word Matrix	544
8.10	Illustration of Latent Semantic Analysis Dimensionality Reduction	544
8.11	Graphical Plate Representation of Latent Dirichlet Allocation	545
8.12	Project Essay Grade (PEG) Model Training Pipeline	554
9.1	A Technology Enhanced U.S. History Item Using Multimedia	591
9.2	PISA 2015 Interactive Science Item in Which Students Use Drag-and-Drop Interactions to Complete a Model of a Fish Farm	592
9.3	PISA 2015 Extended-Interaction Item From the <i>Running in Hot Weather Set</i>	593
9.4	Selected Screens From the NAEP TEL 2014 Scenario-Based Task <i>Bike Lane</i>	595
9.5	A Hidden Markov Model for Estimating Student State Over Time	627
9.6	A Simple Bayesian Network	629
10.1	Example Items Used in Analyses	700
10.2	Descriptive Statistics of Pretest and Posttest Sum Scores by Schools and Conditions	702
10.3	A Path Diagram Depicting the Level 2 (Site-Level) Model	714
10.4	A Path Diagram Depicting the Level 1 (Individual-Level) Model	714
12.1	Typical Standard Setting Workshop Process	823
12.2	The Cognitive and Social Model of Standard Setting	830
13.1	Simplified Sample of a K–12 Results Report	899
13.2	Sample Report Section With Lexile-Based Book Suggestions	900
13.3	A Road Map for Audience-Based Research	903
13.4	Overview of Assessment Contexts for Reporting, per the <i>Standards for Educational and Psychological Testing</i>	904
13.5	Performance-Label Word Cloud	905
13.6	Narrative Patterns for Reporting Data Stories	913
13.7	Simplified Sample Report With Annotations Highlighting Narrative Patterns	914
13.8	Concentric Circles of Interest in Assessment Results	922
13.9	The Hambleton and Zenisky Model for Score Report Development	929

13.10	Key Reporting Recommendations: Overall	938
13.11	Key Reporting Recommendations: Content—Report Introduction and Description	939
13.12	Key Reporting Recommendations: Content—Scores and Performance Levels	939
13.13	Key Reporting Recommendations: Content—Other Performance Indicators	940
13.14	Key Reporting Recommendations: Content—Other	940
13.15	Key Reporting Recommendations: Language	941
13.16	Key Reporting Recommendations: Design	941
13.17	Key Reporting Recommendations: Interpretive Guides and Ancillary Materials	942
15.1	Sample Graph of Curriculum-Based Measurement (CBM) Measures Collected Over 27 Weeks of Instruction, With an Instructional Change Occurring After Week 16	1054
16.1	A Framework for K–12 Test-Based Accountability Purposes and Mechanisms	1116
16.2	Illustrating Mechanisms for Inflation When Test Developers Sample Predictably From Content Domains	1162
17.1	Percentage of 4-Year Institutions by Undergraduate Admission Rates	1208
19.1	Cross-Sectional Comparisons of CORE District Students Scale Scores Across Social-Emotional Learning (SEL) Scales (Top Four Panels) and Achievement Test Scores (Bottom Two Panels)	1364
20.1	International Adult Assessments	1430
20.2	The Virtuous Spiral	1435
21.1	Central Components of the Process of Measurement	1536
21.2	Trends in National Assessment of Educational Progress Reading Comprehension for Grade 4 Students From 1998 to 2022	1542
21.3	Mock Report of Student Results From Smarter Balanced English Language Arts/Literacy Assessment	1543
21.4	Visual Representation of a Sociocultural Theory of Reading Comprehension From the Public Comment Draft of the Reading Framework for the 2026 National Assessment of Educational Progress	1545
21.5	Public Dissemination of Results From National Assessment of Educational Progress 2022 Reading Assessment	1556
21.6	Public Dissemination of National Assessment of Educational Progress Reading Results in the Metric of Percentage <i>Proficient</i> or <i>Advanced</i>	1558

Foreword

Over the 74 years since the publication of the first edition of *Educational Measurement*, this volume has become a signature contribution of the National Council on Measurement in Education (NCME). It is the handbook of our field. Its various chapters provide authoritative definitions of terms, clarify essential concepts, review what we know and have learned, and identify fruitful areas for debate and further study. Whereas the American Council on Education led or partnered with NCME on former editions, this edition is the first that NCME led from conception to delivery.

The NCME Board appointed Dr. Linda L. Cook and Dr. Mary J. Pitoniak as coeditors in 2018. They have steered this volume, its authors, and its reviewers through global disruptions in health, politics, technology, and the central context of this volume, education. In stormy seas, how can we know where we are or where we are going? Drs. Cook and Pitoniak have been intrepid navigators, guiding authors and reviewers through uncharted waters and finding safe harbor to land this volume where it can make its deserved impression on the field. On behalf of NCME, I extend my gratitude and express my admiration for their leadership and accomplishments.

Many people—103 in total—were involved with this volume, some playing multiple roles. They include 52 authors, 63 reviewers, and 19 editorial board members. Like the coeditors, these individuals received no financial compensation for their considerable efforts to develop and improve these chapters. The many hours, days, weeks, months, and years of unpaid work that they contributed reflect their commitment to scholarship itself as a form of public and professional service. NCME, the field, and the public can honor their contributions by reading, engaging, citing, applying, and advancing their work in future research, policy, and practice.

One of these contributors was Stephen (Steve) Ferrara. Steve served as an author of Chapter 12 in this volume, an editorial advisory board member, and a reviewer. Steve also authored a chapter in the fourth edition of this volume and has made innumerable contributions to NCME as an organization, including as a board member and an editor. Steve passed away in 2024 after a prolonged illness. In his honor, Susan Trent and the Ferrara family have made a generous contribution to make this volume freely available to the public. Thanks to their generosity and in Steve's memory, all scholars, students, practitioners, professionals, and citizens have equal access to this signature NCME contribution, regardless of their financial means. Thank you to Susan and the Ferrara family, and thank you, Steve.

On behalf of the NCME community and our field, I extend my gratitude to the coeditors, editorial board members, authors, and reviewers for their generous contributions. The organization is committed to advancing their work to benefit scholarship and society.

Andrew D. Ho
President (2024–2025)
National Council on Measurement in Education

Acknowledgments

An undertaking of this size cannot reach a successful conclusion without the effort and talent of many different people. For the fifth edition of *Educational Measurement* (EM5), we would like to acknowledge first the contributions of the 52 authors who worked diligently to provide the very best exploration and documentation of the most important topics in our field. We would also like to acknowledge the contributions of the 63 reviewers of the 21 chapters appearing in this volume. We were extremely impressed by the thoughtful and insightful comments they provided. The quality of the content of EM5 is reflective of the patience, professionalism, and hard work of these reviewers as well as the responsiveness of the authors to the reviews.

We would like to acknowledge the invaluable advice and guidance we received from the 19 members of the Editorial Advisory Board. Their breadth of talent and experience helped us set the direction for the volume and maintain our goals throughout the process. Each member also served as an additional reviewer of a chapter.

In addition, we would like to acknowledge the National Council on Measurement in Education (NCME) Board for the unending support that they provided throughout the creation of this volume. We would also like to acknowledge the support and guidance of the NCME publications committee.

We are fortunate to have had excellent support and guidance from our colleagues at Oxford University Press: Abby Gross, Sarah Ebel, Hayley Singer, and Katharine Pratt. Their expertise and patience were an immense help to us as we prepared our final drafts of EM5.

We would like to acknowledge the generous support of ETS, which made it possible for the editors to have the time it takes for a project of this magnitude. We would also like to acknowledge the excellent editorial assistance of ETS staff members Jodi Thompson and Kimberly Fryer. Last, we would like to thank two of ETS's officers, Ida Lawrence and Karen Barton, for the many ways they showed their support during the creation of this volume.

Contributor List

In this section, we list and thank the many people involved in conceptualizing, writing, and reviewing this volume.

Author List, by Chapter

1. Linda L. Cook, ETS (Retired)
Mary J. Pitoniak, ETS
2. Brian E. Clauser, NBME (Retired)
Jerome C. Clauser, American Board of Internal Medicine
Amanda L. Clauser, NBME
3. Kadriye Ercikan, ETS
Guillermo Solano-Flores, Stanford University
4. Suzanne Lane, University of Pittsburgh (Emerita)
Scott F. Marion, National Center for the Improvement of Educational Assessment
5. Won-Chan Lee, University of Iowa
Deborah J. Harris, University of Iowa (Retired)
6. Rebecca Zwick, ETS (Retired), University of California (Emerita)
7. Kristen Huff, Curriculum Associates
Paul Nichols, Retired
M. Christina Schneider, Cambium Assessment
8. Mark D. Shermis, Performance Assessment Analytics
Allan S. Cohen, University of Georgia
Jordan M. Wheeler, University of Nebraska–Lincoln
9. Randy E. Bennett, ETS
Michelle LaMar, ETS (Retired)
John Mazzeo, ETS (Retired)
10. Li Cai, University of California, Los Angeles
Mark Hansen, University of California, Los Angeles
Minjeong Jeon, University of California, Los Angeles
Michael C. Edwards, Arizona State University

11. Tim Moses, Buros Center for Testing
12. Steve Ferrara, HumRRO
Susan Davis-Becker, ACS Ventures
Priya Kannan, WestEd
Katherine Reynolds, Boston College
13. April L. Zenisky, University of Massachusetts Amherst
Francis O'Donnell, NBME
Ronald K. Hambleton, University of Massachusetts Amherst
14. Michael C. Rodriguez, University of Minnesota Twin Cities
Martha L. Thurlow, University of Minnesota Twin Cities (Retired)
15. Susan M. Brookhart, Duquesne University (Emerita)
Charlie DePascale, Independent Consultant
16. Andrew D. Ho, Harvard Graduate School of Education
Morgan S. Polikoff, University of Southern California
17. Wayne J. Camara, Law School Admission Council
Ou Lydia Liu, ETS
Krista Mattern, NBME
18. Melissa J. Margolis, NBME
Rebecca S. Lipner, American Board of Internal Medicine
Chad W. Buckendahl, ACS Ventures
19. Patrick C. Kyllonen, ETS
Jiyun Zu, ETS
20. Henry I. Braun, Boston College
Irwin Kirsch, ETS (Retired)
21. Derek C. Briggs, University of Colorado Boulder
Andrew Maul, University of California, Santa Barbara
Joshua A. McGrane, Assessment & Evaluation Research Centre, The University of Melbourne

Reviewer/Collaborator List, by Chapter¹

2. Mark D. Reckase, Michigan State University (Emeritus)
Howard Wainer, Independent Consultant
3. Robert J. Mislevy, University of Maryland (Emeritus)
Richard Duran, University of California, Santa Barbara
4. Michael Kane, ETS (Retired)
Paul Newton, Ofqual

¹ There were no reviewers for Chapter 1. Reviewers are listed in alphabetical order for each chapter. Members of the Editorial Advisory Board also provided reviews, as indicated in the list of its members.

5. Edward Haertel, Stanford University (Emeritus)
Seonghoon Kim, Hanyang University
6. Lorrie Shepard, University of Colorado Boulder
Michael Zieky, ETS (Retired)
7. Madison A. Holzman, Holzman Consulting
Mari Pearlman, Pearlman Education Group
Phoebe Winter, Independent Consultant
8. Tim Davey, ETS (Retired)
Tahereh Firoozi, University of Alberta
Mark Gierl, University of Alberta
Susan Lottridge, Cambium Assessment
9. Ryan Baker, University of Pennsylvania
Denny Way, Retired
10. Daniel Bolt, University of Wisconsin–Madison
Irina Moustaki, London School of Economics
11. Robert L. Brennan, University of Iowa (Emeritus)
Neil J. Dorans, Independent Consultant
Michael J. Kolen, University of Iowa (Emeritus)
12. James Pellegrino, University of Illinois Chicago (Emeritus)
Barbara Plake, University of Nebraska–Lincoln
13. John Hattie, University of Melbourne
Diego Zapata-Rivera, ETS
14. Maria Araceli Ruiz-Primo, Stanford University
Michael Russell, Boston College
15. Dylan Wiliam, UCL Institute of Education (Emeritus)
Mark Wilson, University of California, Berkeley
16. Laura S. Hamilton, National Center for the Improvement of Educational Assessment
Joan Herman, University of California, Los Angeles (Emerita)
17. Jerome Lucido, University of Southern California
Emily Shaw, College Board
18. David Swanson, University of Melbourne
Nancy Tippins, The Nancy T. Tippins Group, LLC
19. Rob Meijer, University of Groningen
Fred Oswald, Rice University
20. Matthias von Davier, Boston College
Hans Wagemaker, Independent Consultant to IEA

21. Neil J. Dorans, Independent Consultant
George Engelhard, The University of Georgia
Mark Wilson, University of California, Berkeley

3. Editorial Advisory Board²

- Robert L. Brennan, University of Iowa (11)
Derek C. Briggs, University of Colorado Boulder (11)
Susan Brookhart, Duquesne University (Emerita) (13)
Kadriye Ercikan, ETS (2)
Steve Ferrara, HumRRO (9)
Ellen Forte, edCount, LLC (12)
Drew H. Gitomer, Rutgers University (15)
Joan Herman, University of California, Los Angeles (Emerita) (4)
Andrew D. Ho, Harvard Graduate School of Education (10)
Michael Kane, ETS (Retired) (4, 14)
Leslie Keng, NBME (18)
Suzanne Lane, University of Pittsburgh (Emerita) (19)
Michael C. Rodriguez, University of Minnesota Twin Cities (16)
Edynn Sato, Sato Education Consulting (7)
Anat Ben-Simon, NITE (Retired) (20)
Martha L. Thurlow, University of Minnesota Twin Cities (Retired) (3)
Frank C. Worrell, University of California, Berkeley (6)
Feifei Ye, RAND (5)
Rebecca Zwick, ETS (Retired), University of California (Emerita) (17)

² Members of the Editorial Advisory Board are listed alphabetically. Members also provided reviews of chapters; chapter numbers are indicated in parentheses.

Part I

Setting the Context

Factors Influencing Education and Educational Measurement

Linda L. Cook

ETS (Retired)

Mary J. Pitoniak

ETS

The past four editions of *Educational Measurement* began by comparing the then-current status of the science of educational measurement with the status of the science during the periods when the previous edition(s) were published. For this edition, we have chosen instead to reflect on the status of society, mostly in the United States, that has influenced both education and educational measurement at the time the chapters in this book were written. We focus particularly on the factors impacting education and educational measurement that we think have been unique influencers during the decade prior to the publication of this book (2014–2024), as compared to the time periods when previous editions were written.

Many of the chapters in this volume were written during one of the most challenging periods in recent U.S. and world history. The outbreak of COVID-19, which began in Wuhan, China, sometime in the fall of 2019, had a profound impact on education and educational measurement as it is practiced in the United States and around the world. As this chapter is being written, the full impact of the pandemic on student learning and other aspects of child development, as well as the impact on society in general, is still being measured and may never be completely known.

Another important factor that is far more complex and difficult to evaluate than the influence of the pandemic on education and educational measurement is the polarized cultural/political climate in the United States, which has intensified the cultural differences impacting what is taught and tested in schools in the early 21st century. Many of the institutions that serve our society, particularly our schools, have experienced changes in policies and practices that reflect conservative cultural preferences, which are sometimes highly controversial.

Finally, and closely connected to the previously mentioned factors, are issues related to social justice. We believe that these issues, which have always been prominent in U.S. society, have intensified and had a more prominent effect on education and educational measurement over the decade prior to the publication of this book (2014–2024) than in previous periods in our history. This well-justified concern about social justice has led to the questioning of many of our institutions and how they serve all members of U.S. society. The increasing awareness of how our society addresses social justice for all our citizens has profound implications for the future of educational measurement.

Although these three factors were not the only factors impacting the science of educational measurement during the past decade, we believe they represent significant societal influences that were not strongly present when the fourth edition of *Educational Measurement* was published in 2006. Some of the chapters in this volume reflect these influences more directly than others and contain explicit descriptions of how the influences affected their topics. Other chapters were influenced by one or more of these factors in some way, but the influence may not be as apparent.

The following section contains a short discussion of the impact of the pandemic, cultural disagreements, and the social justice movement on the recent practice of education and educational measurement in the United States. This discussion is followed by

an overview of each of the five parts in this volume and a short summary of the chapters included in each part. We are hopeful that this material will provide a context for the chapters in this volume, as well as a guide to help readers choose chapters that are most interesting to them or most appropriate to inform their work.

THE PANDEMIC (COVID-19) AND EDUCATION AND EDUCATIONAL MEASUREMENT

The fifth edition of *Educational Measurement* was written and edited during one of the most unique periods in U.S. history that either of the editors has experienced. This period was marked by a worldwide pandemic, COVID-19, which began impacting life in the United States in early 2020. The impact of the pandemic continues to be felt in many ways.

As we write this chapter, for most people, the danger from the disease itself has been reduced to no more than that of a bad cold. However, everyone remembers those terrifying months before a vaccine was developed when young and old alike were succumbing rapidly to the disease.¹

Society as we had experienced it basically stopped functioning. In the early months of the pandemic, businesses closed, schools shut down, residents fled cities, and entertainment centers were shuttered. Sporting events were carried out without an audience. Cardboard cutouts placed in the stadium seats took the place of fans to create the feeling of an audience for those watching the event on television. Sick individuals were isolated and tragically passed away, with their families having only a tablet device or phone to use for their final communications, if they were able to have any communication at all.

We think it is important to understand the fear and despair that gripped U.S. citizens as well as international populations during the early stages of the pandemic. The heartbreak and emotions brought about by the spread of COVID-19 had a lasting effect on the well-being and mental health of the parents, children, and educators who struggled to continue with student learning through the first few years of the pandemic. People lost family members, their jobs, their incomes, and their homes. This loss was not uniformly felt across subpopulations, which exacerbated previously existing societal issues.

Many K–12 schools in the United States coped with the pandemic by rapidly transitioning from an in-person setting to remote instruction offered over the Internet. However, a few schools continued to operate completely in person, while others offered a hybrid approach to classes: 1 or 2 days a week with in-person instruction and the remaining days with remote learning. According to Camara (2020), school closures due to COVID-19 in the spring of 2020 “affected at least 124,000 of the estimated 132,000 U.S. public and private schools and over 56 million students” (p. 10).

As just mentioned, the impact of COVID-19 was not uniform. According to the Center for Reinventing Public Education (2021), “the human losses and financial tolls were more commonly experienced by Black, Hispanic/Latinx, and indigenous students from

lower-income communities” (p. 5). These researchers pointed out that in the spring of 2021, only 47% of high-minority schools had resumed in-person schooling as compared to 61% of low-minority districts.

School closures caused a particular problem for spring state-accountability assessments. Gathering students for assessment in the spring of 2020 was impossible in most situations. Many states simply did not test in the spring of 2020 or even in the fall of that year. Some states experimented with their accountability assessments by administering them individually via the Internet. According to Camara (2020),

In early spring 2020 the vast majority of US colleges and schools closed for the year due to COVID-19. . . . School closures and associated health concerns halted large scale admissions testing and required alternative models such as remote proctoring at home, additional flexibility in test sites . . . and additional testing dates for the fall. . . . It is clear that access to admissions testing has been greatly reduced despite these efforts resulting in extended deadlines in graduate admissions and the wide-scale adoption of test-optional policies in undergraduate admissions. (p. 10)

As we write this chapter, some of the alternative models just referred to have been adopted as a regular part of test administration processes. For example, remotely administered exams have persisted and new ways of protecting the security of these exams have been developed, some more successful than others. In addition, in the spring of 2024, the SAT moved almost completely to online administration, providing more flexibility in the face of school closings in the future.

Researchers will continue to study the impact of the pandemic on student learning for many years. The lack of state assessment data collected during the pandemic years has complicated things considerably. Early indications were for sizeable math and reading drops, exacerbating test score gaps between students in low-poverty compared to those in high-poverty elementary schools (Kuhfeld et al., 2022). However, these researchers were optimistic about some improvements in learning if the school district took advantage of resources from the Elementary and Secondary School Emergency Relief Fund established as part of the Coronavirus Aid, Relief, and Economic Security (CARES) Act in March 2020 (U.S. Department of the Treasury, 2020).

Results from the National Assessment of Educational Progress (NAEP) also provided information about student performance related to the pandemic. In 2022, NAEP results made news across many major platforms:

U.S. students in most states and across almost all demographic groups have experienced troubling setbacks in both math and reading, according to an authoritative national exam released on Monday, offering the most definitive indictment yet of the pandemic’s impact on millions of schoolchildren. (Mervosh & Wu, 2022, para. 1)

However, more recent data are encouraging. In a recent report, *New Data Show How the Pandemic Affected Learning Across Whole Communities* (Education Recovery

Scorecard, 2024), researchers noted that both math and reading scores for elementary and middle school students are showing recovery. The results differ from state to state, with some states showing considerably better recovery than others.

The editors also feel some degree of optimism. We are hopeful that the lessons we have learned from coping with the pandemic, as well as an increased awareness of some of the inequities in our society that were made so apparent as we attempted to address the many COVID-related issues associated with learning and assessment, will result in improvements in how we teach and assess. (The reader is referred to Chapters 3, 6, 9, 11, 16, 17, and 20 for further discussion of the effect of COVID-19 on society, education, and educational measurement.)

THE POLITICAL/CULTURAL CLIMATE AND EDUCATION AND EDUCATIONAL MEASUREMENT

The political/cultural climate in the United States has a long history of affecting both education and educational measurement. A well-known example is the teaching of evolution in public school science classes. The Scopes trial of 1925 is often regarded as a landmark in the battle to teach evolution in American public schools. A high school teacher in Dayton, Tennessee, John T. Scopes, was charged with violating state law by teaching Charles Darwin's theory of evolution. According to "Scopes Trial" (2024, para. 1), "the trial's proceedings helped to bring the scientific evidence for evolution into the public sphere while also stoking the national debate over the veracity of evolution that continues to the present day." In 1987, the Supreme Court ruling in *Edwards v. Aguillard* struck down a Louisiana law requiring schools to teach creation science whenever students learn about evolution: "In issuing its ruling in *Edwards*, the court effectively closed the door on efforts to teach the biblical creation story in public school science classes" (Masci, 2007, para. 1). However, according to a recent article in *Scientific American*, despite the 1987 ruling and the 2013 introduction of the Next Generation Science Standards, there is still reason for concern:

More than one in six high school biology teachers, 17.6 percent, are still presenting creationism as a scientifically credible alternative to evolution. And almost as many high school biology teachers, 15 percent, are still failing to emphasize the broad scientific consensus on evolution, despite its general prevalence in state standards and despite encouragement from professional organizations. (Branch & Reid, 2020, p. 5)

Another example of how cultural differences have affected education in public schools was the concern about socialism or Marxism in the early and mid-20th century. Groups like the American Legion pressured schools to eliminate the use of progressive textbooks that encouraged students to think about economic inequality; later, the John Birch Society raised many of the same criticisms of school curricula (Birnbach, 2021).

Cultural differences and opinions about what should be taught and tested in public schools have been prominent during the drafting of the chapters in this volume, and these differences may be even stronger in the early 21st century than in 1925 when the teaching of evolution was an issue or in the mid-20th century when the concern regarded teaching about economic inequality. A current cultural issue impacting what is taught and tested in schools is *wokeness*, basically understood as liberal views on race and gender. A major concern in many schools has been the teaching of critical race theory.² According to an article in *Education Week*,

since January 2021, 44 states have introduced bills or taken other steps that would restrict teaching critical race theory or limit how teachers can discuss racism and sexism. . . . Eighteen states have imposed these bans and restrictions either through legislation or other avenues. (Schwartz, 2021, para. 5)

States with conservative legislatures, such as Florida, have focused on “scrubbing textbooks.” These states are concerned about textbooks promoting a “leftist agenda” in the classroom (Mervosh & Goldstein, 2023). The Florida Department of Education said in a news release that they rejected 82 of 101 submitted textbooks because of what they considered “inaccurate material, errors and other information that was not aligned with Florida law” (Mervosh & Goldstein, 2023, para. 2). Textbooks are being rejected because they include topics such as critical race theory and social-emotional learning. Examples of topics that are being removed from some textbooks are those focusing on the Black Lives Matter movement³ and the murder of George Floyd⁴ and its impact on society.

Harris (2024) reported on a nationwide surge in book bans affecting memoirs and novels that address experiences of people in the lesbian, gay, bisexual, transgender, and queer (LGBTQ) community or deal with race. The source of the data was a report released by the American Library Association (2023), which described an “explosion” of book bans across the country, resulting in 4,200 challenged titles in 2023. This number represents a 65% increase over the number of challenges raised in 2022. In the years leading up to 2021, when bans began to escalate, the average number of books challenged in a year was about 275.

A highly controversial example of scrubbing textbooks was the state of Florida’s challenge of the pilot version of the College Board’s Advanced Placement (AP) curriculum for African American studies. According to Hartocollis and Fawcett (2023, para. 5), “the College Board’s decision to try to build a curriculum covering one of the most charged subjects in the country—the history of race in America—may have all but guaranteed controversy.” The attack on the AP pilot curriculum began in Florida when the governor announced in January 2024 that he would ban the curriculum. Hartocollis and Fawcett (2023) wrote, “State education officials said it was not historically accurate and violated state law that regulates how race-related issues are taught in public schools.” It should be noted that the curriculum had input from, and was celebrated by, many eminent scholars as an “affirmation of the importance of African American studies” (Hartocollis & Fawcett, 2023). According to these authors,

the study of contemporary topics—including Black Lives Matter, incarceration, queer life and the debate over reparations—is downgraded [in the revised curriculum]. The subjects are no longer part of the exam and are simply offered on a list of options for a required research project. (para. 15)

One of the writers expunged from the AP curriculum was Kimberlé W. Crenshaw, a law professor at Columbia, whose work the university cites as being “foundational in critical race theory” (Hartocollis and Fawcett, 2023, para. 17). Crenshaw said she was disappointed because “she had believed the course would capitalize on a hunger of young students to learn ways of thinking about things like police brutality, mass incarceration and continuing inequalities” (Hartocollis & Fawcett, 2023). Instead, she said, “the very same set of circumstances that presented the need for the course also created the backlash against the content that people don’t like” (Hartocollis & Fawcett, 2023, para. 20).

The changes in the curriculum did not completely satisfy conservatives either. For example, Ilya Shapiro, director of constitutional studies at the Manhattan Institute, criticized the curriculum for omitting conservative or independent Black thinkers like John McWhorter, Shelby Steele, Thomas Sowell, and Justice Clarence Thomas of the Supreme Court (Hartocollis & Fawcett, 2023).

THE SOCIAL JUSTICE MOVEMENT AND EDUCATION AND EDUCATIONAL MEASUREMENT

According to a report from the United Nations (2006), “Social justice may be broadly understood as the fair and compassionate distribution of the fruits of economic growth” (p. 7). The same report that provides this definition also provides a brief global history of social justice. According to this report, the idea of social justice was first noted in Western politics soon after the Industrial Revolution, along with the development of the socialist doctrine. The concept of social justice developed further as a protest of what was then thought to be the capitalist exploitation of labor. The concept of social justice continued to develop and became central to most leftist and centrist political parties around the world by the mid-20th century.

Wikipedia lists over 140 social justice movements that are active as we write this chapter (“List of Social Movements,” 2024). These movements range from the well known, such as the Black Lives Matter movement and the civil rights movement, to the lesser known, such as movements like the sunrise movement⁵ and the slow movement.⁶

Following World War II, the civil rights movement had a very important effect on multiple groups that were protesting for equal rights. These groups included women, Black people, and gay people. One movement that has had considerable impact on current society and education and educational assessment is Black Lives Matter, which we described earlier. Education and educational testing have also been influenced by the LGBTQ social movements.

The gay liberation movement, which began in the 1960s, was founded to encourage gay men and lesbians to engage in radical action to promote gay pride (Bill of Rights Institute, 2024). One of the most noted landmarks in the history of the gay liberation movement was the resistance to the police raid of the Stonewall Inn in New York City that occurred in 1969. This incident was a significant milestone in organizing LGBTQ people to become more active politically and accomplish much for the civil rights of the LGBTQ community (Bill of Rights Institute, 2024).

As we write this chapter, there is a growing global pro-Palestinian protest that has risen out of the Israel– Hamas war. The protests in the United States are located mostly on college campuses. The protesters are demanding an end to Israel’s war in Gaza, which at the time of this writing had killed more than 34,000 Palestinians. In addition, they are making other demands: for example, that universities divest themselves from financial ties with Israel or from companies that may profit from the war in Gaza. The protests have disrupted college and university classes and graduation ceremonies and led to a number of arrests.

In addition, the protests have sparked a debate over freedom of speech on campuses that has resulted in the resignation of two presidents of prominent universities. Four university presidents have been asked by Congress to testify about “rampant antisemitism” on their campuses. The four presidents were from the University of Pennsylvania, Harvard, the Massachusetts Institute of Technology, and Columbia. At the time of this writing, two of the four presidents have resigned from their university positions. The president of Columbia, Minouche Shafik, warned, “If colleges and universities cannot better define the boundaries between free speech and discrimination, government will move to fill the gap, and in ways that do not necessarily protect academic freedom” (Forgash, 2024, para. 7).

At present, those involved in educational assessment have responded mostly to social justice movements focused on groups such as people of color and those associated with gender, disability, or sexual orientation. Often “equity indicators” are used to discuss and evaluate the fairness and equity of education and assessment for these groups (Edley et al., 2019). In contrast, Vo (2021) used a social justice frame rather than an equity frame to examine the design and interpretation of opportunity-to-learn indicator systems. She argued that an equity frame “maintains a traditional view of opportunity-to-learn, OTL, that does not explicitly consider the complexities of how knowledge is acquired or measured” (p. 1). She contrasted an equity frame with a social justice frame using an example from Edley et al. (2019). She noted that as an indicator of equity in a school setting, the authors described using “disparities in engagement in schooling” (Vo, 2021, p. 25), which they quantified as 15 or more unexcused days absent from school in 1 year. Vo pointed out that this is also the most common school quality and student success indicator used in the Every Student Succeeds Act (U.S. Department of Education, 2015).

Vo (2021) continued by postulating that using an equity frame “could lead to interpreting and labeling a student as ‘disengaged’ but neglecting to understand the factors

why students might not be attending school in the first place” (p. 2). She made the point that a social justice frame would “challenge the definition of ‘student engagement’” (p. 2) and would examine the idea of engagement by looking at the cultural experiences of individuals who are Black, Indigenous, and people of color. As an example, she described a student who is self-employed, living with a single parent, and helping to provide child care and income for the family. If the school were to apply a social justice frame rather than an equity frame to this student’s situation, they would look at the sociocultural context. By doing this, the school might conclude that the student’s lack of engagement is a result of his family and community experiences. Using this frame, the school might be able to provide support to the student, both in and outside school, that would result in improved school engagement.

Michael Russell (2024), in his book addressing systemic racism and educational measurement, began his chapter on the “White racial frame” with a quote taken from Elias and Feagin (2016):

White racial framing involves the explanation and construction of social reality from the perspective of dominant whites, one normally steeped in Eurocentrism, thereby creating a long-lasting white racial frame. This white racial frame is a meta-structure that develops and reinforces ideas, actions, networks, institutions, and social structures according to the views and practical racial-group interest of white people. (p. 7)

Russell devoted an entire chapter in his book to exploring the White racial frame and how it influences educational measurement. He concluded this chapter by making the point that “while the racist beliefs that dominated the thinking of the field’s pioneers no longer operate in an overt manner, the framing of questions, methods, and narratives were carried forward to influence educational measurement as it operates today” (p. 126).

In his book, Russell (2024) referred frequently to the work of Jennifer Randall. Randall, the founding president of the Center for Measurement Justice, focuses her work on “[disrupting] white supremacist, racist logics in assessment through justice-oriented practices that are explicitly and unapologetically antiracist” (Randall, 2021). Randall (2021) discussed an “antiracist framework for construct definition and representation via an approach that actively disrupts current white supremacist social, economic, and political ideologies that reproduce racial privilege in society through the promotion of white, racist, hegemonies” (p. 82). In this article, she argued that the current practice of “committing to neutrality (or an absence of racism) is, in effect, sustaining white supremacist logics and continues to actively harm minoritized students” (Randall, 2021, p. 83). She went on to say that “the foundational assessment design’s commitment to neutrality—operationalized by our fierce eradication of anything that could be considered ‘construct-irrelevant variance’—is, in effect, sustaining the inherent whiteness as superior bias of our assessment design practices” (Randall, 2021, p. 84). She continued this argument in a discussion of assessment validation

(Randall et al., 2022). Randall and her colleagues made the point that “the validation process, arguably the most important process in educational measurement, has served as one of measurement’s primary weapons of racist oppression and marginalization” (Randall et al., 2022, p. 171). The authors went on to propose a justice-oriented, anti-racist validity framework that is “built on foundational theories and practices related to justice oriented perspectives, critical race theory, and antiracist assessment” (Randall et al., 2022, p. 171). Randall et al. (2022) made a critical observation that “what is construct-relevant behavior is dependent on the eye of the beholder” and pointed out that what validity theory has been missing is a set of questions that “ground and initiate assessment design and validation” (p. 173). They note that Mislevy (2018) argued that “tasks, performances, and criterion situations need to be examined through the lens of targeted LCS (linguistic, cultural, and substantive) patterns and practices” (Randall et al., 2022, p. 173). For example, Randall et al. asked “whose LCS patterns are being privileged by an assessment, whose LCS patterns are being devalued, omitted, suppressed, or marginalized?” (p. 173).

References to and discussions of the relationship between social justice and assessment are also found in Chapters 2, 3, 4, 6, 14, 15, and 17.

OVERVIEW OF THE VOLUME

Part I: “Setting the Context”

Part I consists of three chapters. In this chapter, Chapter 1, “Factors Influencing Education and Educational Measurement,” we attempt to provide some societal context for the 20 other chapters that appear in this volume. The other 2 chapters in this part are Chapter 2, “The History of Educational Measurement,” and Chapter 3, “The Socio-cultural Context of Educational Assessment.” The chapters in Part I provide a more in-depth discussion of the historical and societal factors influencing education and educational measurement in general. Chapter 2 provides a comprehensive discussion of the history of the science of educational measurement and Chapter 3 includes a thoughtful discussion of how educational measurement influences and is influenced by the society it serves.

Brian E. Clauser, Jerome C. Clauser, and Amanda L. Clauser, the authors of Chapter 2, provide a quote by Howard Wainer (2005, p. 118), attributed to Aristotle: “We understand best those things we see grow from their very beginnings.” They go on to say that “if we wish to understand the field of educational measurement, we would be well served by understanding how it developed” (p. 92). They point out that the history of educational measurement is not a simple one, but rather multiple related histories. To accommodate this complex history, they have separated their chapter into eight sections. They cover topics such as early foundations, intelligence testing, the impact of the U.S. government on educational testing practice, test theory, scaling, item response theory, and the use of tests for selection and accountability.

An important discussion in Chapter 2 that helps us understand concerns about testing and social justice in the early 21st century focuses on the fact that many of the early, and most prominent, thinkers in statistics, psychometrics, and educational measurement were closely aligned with the eugenics movement. They point out that

viewing measurement and testing in the context of an early history that is closely tied to intelligence testing and eugenics is likely to help us understand the skepticism of and resistance to testing that has been present since the 1920s. (p. 35)

Another important point the authors make related to the topic of social justice is to

suggest that the emphasis that Cronbach (1971, 1989b) and Messick (1975, 1989) placed on fairness and consequences as part of test validity can be accounted for (at least in part) by the social context created by the civil rights movement and the war on poverty. (pp. 34–35)

Chapter 2 also includes an insightful discussion of the political forces that have shaped testing over the years. As we discussed in the previous section, the political climate in the United States, as in other parts of the world, is key to the direction taken by our social institutions such as education. The authors of the chapter point out that as early as 1860, the U.S. government made important decisions about what to support and not support in testing. They point to the civil service commission and the army testing program as examples. Readers, particularly those interested in the nexus of policy, social justice, and assessment, will find many of the insights presented in the chapter timely and informative.

Kadriye Ercikan and Guillermo Solano-Flores, the authors of Chapter 3, “The Sociocultural Context of Educational Assessment,” begin their chapter by stating that their goal is to make clear that education and the sociocultural context within which it takes place have a bidirectional relationship. They highlight this relationship as it manifests in different ways for measurement professionals, researchers, decision-makers, and practitioners. They point out that

contexts determined by conditions such as legislation, educational policy and practices, technological advances, social inequities, and unexpected events such as the COVID-19 pandemic affect assessment practices . . . In turn, assessment impacts education policy, teaching practices, and the learning, career, and life opportunities of individuals. (p. 116)

The authors discuss three notions that drive their reasoning about the bidirectional relationship of assessment and sociocultural contexts: first, culture and language influence how test takers engage with tests, process information presented on tests, and formulate solutions and responses to test tasks; and second, because tests and testing are social products, they are a reflection of the societies in which they originate; test content is a reflection of the kinds of knowledge, reasoning, and skills valued in a society. They make the important point that “tests tend to be written in a standard, formal variety of the language in which they are administered This type of language usage is not common among certain test takers—and especially so for test takers from

a non-native language background in the language of a test” (p. 117). The third notion they discuss is that assessment has different meanings and roles for different cultures. They point out that

views of tests and testing and the consequences that tests and performance on tests have on people’s lives vary across societies. This creates differences across societies in the ways in which individuals prepare for taking tests, how they interpret test items and their demands, the level of effort they put into responding to tests, and the pace with which they solve problems during a test. (pp. 117–118)

The editors of this volume believe that it is important for those who practice educational measurement to have a thorough understanding of how the science of educational measurement is influenced by history as well as by past and present sociocultural conditions. It is our intention that the chapters found in Part I of this volume provide that opportunity for our readers.

Part II: “Foundational Concepts”

Three chapters appear in Part II; they focus on the foundational topics of reliability, validity, and fairness. It is important to point out that the fifth edition of *Educational Measurement* contains two fairness chapters. The chapters differ in their focus, and because of these differences, they are placed in different parts of the volume. Chapter 6, “Fairness in Educational Measurement: Theory and Concepts,” which appears in Part II, focuses on fairness as a broad foundational concept underlying all aspects of assessment. The second chapter on fairness, Chapter 14, “Realizing Fairness Through Accessibility for All Test Takers and for Specific Groups,” found in Part III, focuses more on recognizing the diversity of test-taking populations and promoting fairness for all test takers through universally designed accessible assessments.

Chapter 4, “Validity and Validation,” authored by Suzanne Lane and Scott F. Marion, is the first chapter in Part II. The chapter expands the argument-based approach to validity discussed by theorists such as Cronbach (1988), Kane (2013), and Shepard (1993). The authors provide an excellent review of how validity, as described in previous chapters of *Educational Measurement*, has evolved. For example, they refer to Kane’s (2006) presentation of an argument-based approach to validity found in the fourth edition of *Educational Measurement* and expand the discussion of the importance consequences have for validation and the validity process. They illustrate this discussion by tying the importance of interpretation and use arguments (IUAs) and validity arguments to different testing programs, including K–12 assessments and credentialing exams.

Chapter 4 also contains a section on culturally responsive and situated assessment practices that includes a discussion of providing meaningful and equitable score interpretations and uses for individuals from diverse cultural backgrounds at a time of “heightened recognition of the long-standing systemic inequities in our assessment and education systems” (p. 194). The authors close their chapter by saying,

We support the call for a justice-oriented framework for conceptualizing constructs and identifying construct-irrelevant sources of variance to better promote diverse cultural and linguistic ways of knowing and doing. . . . Similarly, we must ensure that the design of the IUA and validity argument addresses any inherent biases so these arguments are designed to better ensure valid score inferences and test use for traditionally marginalized cultural and linguistic groups. (p. 261)

The second chapter in Part II is Chapter 5, “Reliability in Educational Measurement,” authored by Won-Chan Lee and Deborah J. Harris. The authors focus on the concept, quantification, and estimation of reliability and errors of measurement and consider these factors primarily under three measurement theories: classical test theory (CTT), generalizability theory (GT), and item response theory (IRT). Reliability is considered based on the concept of *replication*, and a conceptualization estimation framework is suggested for reliability analysis.

The sections in Chapter 5 that focus on CTT and GT build on treatments in chapters appearing in previous versions of *Educational Measurement*. The section on IRT consolidates and expands the work around reliability in IRT, which was not present in previous editions. Special emphasis is placed on conditional standard errors of measurement, with a dedicated section covering their various estimators. The reliability of classification consistency/accuracy is also discussed.

Additionally, Chapter 5 reviews a series of issues related to how we address reliability. These include how to use structural equation models to determine the appropriate precision coefficient when dimensionality is an issue, how speededness may affect reliability in computer adaptive testing, and how reliability is estimated when automated scoring is used with test-taker responses.

The chapter concludes by noting that

since the fourth edition of *Educational Measurement*, there has been relatively little literature published on the integration of IRT, GT, and CTT. . . . Hopefully, this chapter is a meaningful response to that goal, though there is still plenty of room for additional progress. (p. 368)

The third and final chapter in Part II is Chapter 6, “Fairness in Educational Measurement: Theory and Concepts.” Rebecca Zwick begins the chapter by providing a brief context for the fairness discussion that follows. She creates this context by writing about some current factors that are influencing how we think about fairness in the United States in the early 21st century. She discusses the pandemic, which led to the widespread elimination of admissions testing requirements and to large-scale remote testing in other settings. In addition, she points out the increasing awareness of systemic racism in the United States that, in turn, has brought focus on “the role that test scores can have in limiting educational and occupational opportunities” (p. 384). She also provides a historical context with a discussion of the imperial civil service exams that began in China 1,500 years ago. She draws an interesting parallel between the fairness

and validity issues faced by these early exams and the issues associated with many of the assessments we administer today.

Chapter 6 consists of six sections. The first section is historical and sets the context for the following five sections. The second section considers the relationships among the concepts of test fairness, validity, and bias. The third section focuses on the legal perspectives of test fairness and includes a discussion of four landmark cases involving test construction practices, test disclosure policies, the annotation of test scores resulting from nonstandard administrations, and the equitability of college admissions tests. The fourth section presents a statistical perspective on fairness that describes different types of analyses that are currently used to evaluate fairness. The fifth section describes some psychometric models that have been proposed to improve the fairness of selection procedures, and the sixth section includes the chapter summary and conclusion.

In the summary and conclusion section, the author makes it clear that she will not provide a single simple definition of fairness. She points out that the never-ending debate about college admissions, discussed in this chapter, “is an illustration of the fact that there will never be consensus on the definition of test fairness” (p. 426). She does, however, provide a very cogent set of general fairness principles. She suggests, as do the authors of Chapter 4, the first chapter in Part II, that fairness be considered a component of a larger validity argument that is being constructed for an assessment.

Part III: “Developing, Administering, and Scoring Assessments”

There are eight chapters in Part III: “Developing, Administering, and Scoring Assessments.” Some of the chapters in this part—Chapters 7, 8, and 11–14—describe important operations carried out for an assessment, that is, procedures and practices that must take place to design, develop, and administer an assessment and those required to report reliable, fair, and valid inferences from assessment results. These chapters are updates of previous chapters appearing in *Educational Measurement* and include new theories, changes in practice, and technological advances. The two remaining chapters in this part, Chapters 9 and 10, focus on theory and methods that are designed to support the operations described in the other chapters in Part III. These two chapters present new concepts and ideas that are not found in previous versions of *Educational Measurement*.

In the first chapter in this section, Chapter 7, “Designing and Developing Educational Assessments,” Kristen Huff, Paul Nichols, and Christina Schneider focus on large-scale interim and summative assessments. They discuss the evolving complexity of educational assessments and make the point that “contemporary educational testing is faced with additional complex challenges that routine design practices simply cannot meet” (p. 442). The authors group these challenges into four categories: the complexity of target constructs, using assessment results for multiple purposes, assessment quality, and accessibility for increasingly diverse student populations. They argue that assessments designed with conventional approaches do not adequately address these challenges.

They point out that conventional approaches to assessment design and development are fragmented and that there is a need for these processes to be integrated and supplemented. The authors offer principled assessment design as an approach that addresses these issues. They describe the procedures and processes used in the approach and give many practical examples of its application.

The second chapter in Part III is Chapter 8, “Test Administration and Scoring.” Mark D. Shermis, Allan Cohen, and Jordan Wheeler begin the chapter by emphasizing the connection between test administration, scoring, and validity. They point out that “ultimately, test administration and scoring procedures should support the intended interpretations of test scores. When test administration or scoring problems occur, they can undermine the validity of score inferences” (p. 514). The chapter focuses on current and emerging practices in test administration and item-level scoring. The authors discuss ways to maintain the integrity of the test administration through test security. They also talk about human scoring and machine-based scoring built on human models or artificial intelligence. The chapter provides examples of machine-based scoring systems, as well as new technologies for determining text and speech content. The chapter ends with a discussion of research conducted on machine scoring systems and the importance of providing evidence of the appropriateness of the interpretations of the scores produced by these systems.

Chapter 9, “Technology-Based Assessment: Validity, Modeling, and Analysis Issues,” is the third chapter in Part III. Randy E. Bennett, Michelle LaMar, and John Mazzeo define technology-based assessment as “a measurement used for decision-making primarily in education, but also in the workplace, that employs digital computing in most, if not all, aspects of its creation, delivery, presentation, scoring, or reporting” (p. 582). The chapter’s first main section concerns assessments designed to support consequential decision-making purposes, such as NAEP, the GRE General Test, and the Law School Admissions Test (LSAT). The authors discuss four kinds of innovative assessment designs used in such assessments—technology-enhanced items, extended interaction items, scenario-based tasks, and simulation-based performance tasks—and explore some of the validity issues raised by these innovations. The chapter’s second major section focuses on assessments employed for in-the-moment instructional decisions or for describing what students know and can do so that near-term instructional next steps can be taken. The last main section explores the idea of combining these divergent purposes in the same assessment.

The fourth chapter in Part III is Chapter 10, “Modeling Data From Educational Assessments.” Li Cai, Mark Hansen, Minjeong Jeon, and Michael C. Edwards set the context for the chapter by making the points that “data from educational assessments can take on a variety of forms and be collected for a variety of purposes. The models and modeling required to infer meaning from assessment data are also varied.” (p. 656). The authors describe various sources and features of educational assessment data, discuss some fundamental concepts pertinent to statistical modeling, and introduce a general framework that places modeling within a larger validation process. They

present two examples. In the first example, they “take a deep dive” into a complex data set collected from a randomized, controlled multisite trial of a math intervention with pre and post measures. This example illustrates issues around model building, selection, and fit appraisal through the use of a realistically complex scenario. In the second example, the authors use data from a recently developed educational assessment program to highlight the need for advanced statistical modeling in a context that is greatly constrained by legal requirements and practical considerations.

Chapter 11, “Scaling, Equating, and Linking,” notes that it would be difficult to over-emphasize how important reported scores are for a large-scale assessment program. Therefore, “as a testing program’s most visible and widely used products . . . it is essential to understand and document how reported scores are produced, maintained, and, at times, related to the scores reported by other testing programs” (p. 736), material that this chapter covers thoroughly. In the second section of the chapter, Tim Moses updates the discussions of reporting scales found in the previous four volumes of *Educational Measurement*, focusing on approaches to setting scales for large-scale testing programs. The third section covers equating approaches for alternative versions of the same test, and the fourth section describes approaches for relating scores and scales of different tests. The author discusses the relationship between the scaling and equating operations presented in the second through fourth sections and other testing program operations such as test development, scoring, and administration.

The author of Chapter 11 points out that recent trends in testing have created challenges in establishing and maintaining score scales of educational tests. One important challenge discussed in this chapter is the impact of the COVID-19 pandemic, which has resulted in an increase in alternative online administrations with mixed effects on reported scores.

A particularly timely discussion in Chapter 11 focuses on the challenge of addressing increasingly diverse testing populations that has, in turn, led to calls to develop and administer tests under less standardized conditions (Sireci, 2020), including ways that are adapted to test takers’ sociocultural backgrounds (Mislevy, 2018). The author recommends that, in part, these challenges be addressed by testing organizations being forthcoming in communicating the scaling and linking procedures they use, the assumptions they are making, and the extent to which comparability of scores is supported.

Chapter 12, “Standard Setting: A Cognitive and Social Model,” is the sixth chapter in Part III. Steve Ferrara, Susan Davis-Becker, Priya Kannan, and Katherine Reynolds begin the chapter by providing a historical overview of standard setting, highlighting the evolution of methods and policies around setting cut scores. The rest of the chapter focuses on the concept of panelists’ cognitive judgmental task as the kernel of various standard-setting methods. The authors draw on findings in judgment and decision-making research literature and use this information to “consider what may be going on in the minds of standard-setting panelists when they participate in standard-setting workshops and make judgments about cut scores” (p. 823).

To support this framework, the authors of Chapter 12 provide a proposed cognitive and social model of standard setting, a taxonomy of standard-setting methods defined by the underlying cognitive-judgmental tasks, and a comprehensive research agenda that raises various questions for scholars, including the influences of the facilitator on the panelists, interpanelist interactions and group dynamics, and a consideration of the larger sociopolitical context of standard setting. The authors provide suggestions for parsing these sociopolitical influences into those that may be desired and undesired, as well as strategies for designing and leading standard-setting activities to minimize the influence of undesired influences.

The seventh chapter in Part III is Chapter 13, “Reporting Scores and Other Results.” The purpose of this chapter is to “provide readers with a broad theoretical context for understanding reporting and reporting choices and to draw on the growing body of research on reporting test results that is relevant to practical problems” (p. 896). April L. Zenisky, Francis O’Donnell, and Ronald K. Hambleton make the point that when ordinary people think about tests, they are not thinking about models or theories; rather, they are thinking about test scores: What do they mean? People want to know whether they passed the test or whether their child passed the test. The score report is what answers these questions.

The authors of Chapter 13 point out that the topic of reporting test results cuts across the fields of psychometrics, marketing, communications, graphic design, cognitive science, and information processing. They introduce a seven-step model for reporting based on a synthesis of the work done by Hambleton and Zenisky (2013). The chapter includes insightful discussions of challenges and opportunities for score reporting such as next-generation assessments, subscore reporting, technology-based reporting, advances in data visualization, and situations where reporting formative interpretations are prioritized.

Chapter 14, “Realizing Fairness Through Accessibility for All Test Takers and for Specific Groups,” is the eighth and final chapter included in Part III and is authored by Michael C. Rodriguez and Martha L. Thurlow. This chapter is the second of two chapters in this volume that focus on fairness in testing. The first chapter, Chapter 6, appears in Part II, “Foundational Concepts,” and has a broader focus on fairness, emphasizing the importance of documenting evidence to support a combined valid/fair interpretation of test scores. The purpose of Chapter 14 is more focused and, to quote the authors, “is to apply the theory and concepts of a fairness argument to all test-taking populations, both in general and for specific groups of test takers” (p. 960).

The authors of Chapter 14 focus on two critical concepts: accessibility and an expanded version of universal design for assessment that includes universal design for learning. They make the point that fairness challenges are less about the characteristics of the test taker and more about the quality, inclusiveness, and responsiveness of test design and implementation. The chapter includes multiple examples that focus on fairness implications for both practice and research.

A final section in Chapter 14 addresses the “promise of culturally and linguistically responsive assessment to provide accessibility and achieve greater test

fairness” (p. 961). This section includes topics such as the role that cultural and linguistic backgrounds play in testing, CLR construct and domain definitions, test and item specifications that support CLR and assessment for social justice.

Part IV: “Specific Applications of Assessment”

Part IV, “Specific Applications of Assessment,” contains six chapters that focus on current issues in large-scale national and international assessments, licensing and certification, and social-emotional skills. A broad range of topics are covered in Part IV, including accountability, admissions, selection, placement and outcomes assessment, and assessments to inform instruction and learning. The chapters in this part are intended to appeal to a broad audience of educational measurement professionals.

The first chapter in Part IV is Chapter 15, “Assessment to Inform Teaching and Learning.” Susan M. Brookhart and Charles A. DePascale begin the chapter by making the point that in recent years, “expectations have risen that all assessments—not just those designed to be used formatively—should inform teaching and learning” (p. 1018). They describe and evaluate many of the assessments that stakeholders use to inform teaching and learning, both those designed for this purpose and those that have been designed for other purposes.

Eight general types of assessment are explored in Chapter 15: classroom formative assessment, interim assessment, common formative assessment, grading, state summative assessment, curriculum-based measurement, student learning objectives, and assessment based on learning progressions. The authors point out that these types of assessment vary in the nature of the measurement information they provide, their effectiveness, their quality, and their assessment literacy requirements. They emphasize that there is a public belief that any information that is available about student achievement should be used to inform teaching and learning and that this belief puts pressure on vendors and districts to produce products such as subscores and interim assessments.

The authors of Chapter 15 point out that the demand for information to inform teaching and learning requires that assessments are designed to be more student specific and to produce valid and reliable information about individual learners as well as groups. In addition, this information needs to be instructionally sensitive and situation specific so that it is relevant to, and can serve, a learning community, most often a classroom. They acknowledge that assessments need to continue to serve institutional, accountability, or administrative decision-making and produce reliable and valid data for groups within units such as schools, districts, and states. However, the authors suggest that these latter purposes are beginning to be overtaken by the formative purposes of assessment. They point out that this is happening in both effective and ineffective ways and that the field of educational measurement needs to understand this and work to support quality assessment in the face of increasing demands for assessment that supports teaching and learning.

Chapter 16, “Test-Based Accountability in K–12 Education,” is the second chapter in Part IV and is authored by Andrew D. Ho and Morgan S. Polikoff. It is important to note that the first three editions of *Educational Measurement* did not contain a chapter on accountability. The first chapter on accountability, which appeared in the fourth edition, was authored by Daniel M. Koretz and Laura S. Hamilton (2006) and provided a history of accountability in the United States from the 1980s through the No Child Left Behind Act of 2001. Chapter 16 extends this history through the following three national accountability policy periods.⁷ The authors provide a thoughtful discussion of how each of these periods brought its own set of measurement challenges.

Chapter 16 includes 10 sections that focus largely on assessment for accountability in the United States, including NAEP. The authors note that in contrast to the users of individual student scores, “accountability systems often incorporate many actors, and many derived scores, at many levels of an educational system. Who is holding whom accountable? By what mechanism? Using which scores? For what purpose?” (p. 1114). Issues addressed in the chapter include the historical development of K–12 test-based accountability, public accountability as reflected in NAEP, student and teacher accountability, and the impact of accountability policies on K–12 education. Also addressed are derived accountability scores, including growth models; evaluation of the alignment of assessments for accountability; validation of test-based inferences in accountability systems; accountability testing for special populations; and fairness and equity in accountability testing.

The chapter extends the framework and principles developed by Koretz and Hamilton (2006) and Koretz (2008) in relation to the validity of evidence collected under low-stakes administration conditions for accountability to the wide range of derived scores common in modern accountability systems. All topics in this chapter are covered in a clear and thorough nontechnical manner that is very accessible to a broad audience. The discussion of derived scores is particularly notable, with information presented in a nontechnical manner that provides a very accessible means for readers to gain insight into these important models.

The third chapter in Part IV is Chapter 17, “Assessment for Admission, Placement, and Outcomes in Higher Education,” authored by Wayne J. Camara, Lydia Liu, and Krista Mattern. The chapter describes the history, use, and design of many of the tests employed for admissions, placement, and outcomes assessment in the United States in the early 21st century. The authors also discuss the research on the validity, fairness, and reliability of these tests.

In Chapter 17, the authors point out that the role of standardized tests in U.S. higher education has undergone profound changes in the past few years as a result of COVID-19 and the social justice movement, the continued trend toward test-optional policies, and renewed challenges to affirmative action.

They refer to this situation as “a near-perfect storm, which further fueled alternative admissions processes without test scores” (p. 1194).

Chapter 17 includes an extensive discussion of the impact of both COVID-19 and the social justice movement on admissions testing and on higher education in general. In addition, the authors discuss some of the current legal challenges involving admissions testing, including the Varsity Blues case⁸ that resulted in the prosecution and conviction of an admissions consultant and several parents for using fraudulent means to gain advantage in the college admissions process for their children.

The authors conclude Chapter 17 by acknowledging that the use of tests in higher education, particularly for college admissions, remains one of the most controversial test uses in the United States. They point out that the accumulated evidence of validity, reliability, and fairness of the use and interpretation of test scores often has little positive persuasive effect on the general public. They end by saying that

in the next few years, we anticipate the release of additional research, which can better help us evaluate the efficacy and utility of admissions testing, and we hope such research will also address the efficacy and consequences of admissions practices and decisions made based on other factors. (p. 1243)

The fourth chapter in Part IV is Chapter 18, “Assessment for Licensing and Certification,” authored by Melissa J. Margolis, Rebecca S. Lipner, and Chad W. Buckendahl. This chapter includes sections that focus on a historical perspective for licensing and certification, fairness and legal issues, test development, reliability, validity, and setting standards, and it closes with a discussion of the future of assessment for licensing and certification.

The authors begin Chapter 18 by reviewing a number of terms—regulation, credentialing, licensure, certification—that are used to refer to different aspects of occupational oversight intended to protect practitioners, the profession, and the public from harm. They describe each of the terms and explain the nuances that contribute to their meaning. The authors of Chapter 18 point out that licensing and certification tests have differences that set them apart from tests used for other applications. They explain that, in part, these differences result from the higher stakes associated with licensing and certification examinations and the risks associated with granting credentials to unqualified candidates, which could pose serious dangers to the public. This is compared to the risks associated with creating barriers to employment and thus limiting the public’s access to services that could be provided by competent practitioners.

The authors describe the history of licensing and certification assessments and how perspectives and practice in this area have changed over time. They review legal considerations in licensing and certification contexts, particularly with respect to the ability to challenge testing practices and outcomes, and explore fairness considerations with respect to factors such as testing accommodations.

The majority of Chapter 18 is devoted to investigating the importance of fundamental assessment considerations in licensing and certification contexts. Some of the issues discussed relate to test development, including a discussion of “the importance of clearly

specifying the purpose of the test and focusing on test development as a data-driven process informed by an understanding of the knowledge and skills that one must demonstrate in the context relevant to the specific credential” (p. 1269). The topics of reliability, validity, and standard setting also are discussed, with a focus on factors that are particularly important in credentialing contexts, including classification accuracy, extrapolation from test performance to performance in practice, and procedural standard-setting considerations in credentialing contexts. The chapter concludes with a discussion of factors that have influenced change in credentialing assessment. These factors, comprising technological advancements, the rise of competency-based education, the move to integrate learning and assessment, and the rise of antitest and anti-regulation sentiments, have catalyzed recent change.

Chapter 19, “Assessment of Social-Emotional, Character, Behavioral, and Intrapersonal and Interpersonal Skills,” is the fifth chapter in Part IV. Patrick C. Kyllonen and Jiyun Zu begin by stating that studies suggest that measures of noncognitive skills are as predictive of educational, workforce, and everyday life outcomes as measures of cognitive skills are. The authors review what is known about these skills and how best to measure them. They discuss frameworks for key skills and uses of noncognitive skills measures for both high- and low-stakes tests. The authors review the predominant measurement approaches: self- and other-report ratings, anchoring methods, forced-choice and ranking methods, situational judgment tests, and games and simulations.

The authors of Chapter 19 point out that one roadblock to the adoption of noncognitive measures in schools is the belief that personality traits are immutable. They disagree with this characterization and argue that research shows that personal qualities such as work ethic and social skills can be changed with good teaching and appropriate support. They state that “what has largely distinguished noncognitive skills is not the nature of the constructs per se as much as it is the measurement method—cognitive skills tend to be measured with tests and noncognitive skills with ratings” (p. 1391). They conclude by noting that the level of interest in using noncognitive measures in both education and workplace settings “has outpaced the field’s ability to provide adequate measures of those skills” (p. 1392) and that “it is essential that the educational measurement community meets that interest [in noncognitive measures] with appropriate tools” (p. 1392).

The final chapter in Part IV is Chapter 20, “Large-Scale Assessments in International Contexts.” Henry Braun and Irwin Kirsch begin by defining international large-scale assessments (ILSAs) as group score assessments that typically focus on in-school K–12 students or adults 16–65 years of age living in households. These surveys assess and compare the knowledge, skills, experiences, behaviors, and attitudes of populations in multiple jurisdictions and languages. Since their inception, these assessments have experienced remarkable growth in participation and salience. The authors attribute this growth to perceptions of increased utility of the assessment results by policy makers and key stakeholders, resulting from the rapid changes taking place in the early 21st century. There is a continuing need

to respond to these changes with regard to the constructs being assessed and how best to measure them. This includes the introduction of new constructs, as well as the modification of existing constructs in both the cognitive instruments and the background questionnaires. The authors refer to these ongoing enhancements as a virtuous spiral.

Chapter 20 presents a historical context for ILSAs, as well as a framework that provides developers and users of ILSAs with a set of design criteria that can be used to evaluate the potential utility of the assessments. This framework, originally developed by Samuel Messick, is adapted by the authors. The chapter includes a discussion of new technologies and how they impact the management, development, and implementation of ILSAs, along with advances in methodology that impact scaling and data analysis. In addition, there is a discussion of a range of issues, including ways to evaluate the impact of ILSAs, some key technical challenges that must be addressed in future cycles, and political challenges that the growth of these assessments have raised. The authors conclude that

although it is certain that ILSAs will continue to face a number of challenges, we believe that the path along the virtuous spiral will continue to provide innovations that will enable ILSAs to address new and more complex questions. We expect that answers to these questions will lead to continued appreciation for the utility for such surveys. (p. 1491)

Part V: "Measurement as a Science"

Part V, "Measurement as a Science," contains only Chapter 21, "On the Nature of Measurement," authored by Derek C. Briggs, Andrew Maul, and Joshua A. McGrane. The purpose of Part V is to consider the practice and process of educational measurement as a special case practiced in the broader context of measurement in the psychological and physical sciences. In this final part of the volume, different perspectives from the philosophy of science are described and the status of measurement in education is considered from these perspectives.

The authors engage with the question of how educational measurement can be understood with respect to measurement more broadly and whether and how instances of educational measurement earn this claimed trustworthiness. In the second section of the chapter, they trace some key historical developments that help to explain how the concepts and terminology of measurement came to be invoked in educational and psychological contexts. In the third section, they argue that "whether one is speaking of measurement in an educational context, in psychology, or in the physical sciences, the activity will involve a common set of elements that serve to justify its claimed trustworthiness" (p. 1515)

The authors of Chapter 21 introduce a framework for understanding measurement with respect to four central components: theory, instrumentation, scales and units, and modeling. They illustrate and apply the framework briefly to the physical science context of measuring temperature and then more extensively to the educational

context of measuring reading comprehension. Although the central components of measurement are present in each context, there are unique challenges in education. In the fourth section, they take a closer look at three of these challenges that are likely to remain areas for discussion and debate in the future: (a) the meaning of validity and how it should inform measurement practices, (b) evolving conceptualizations of measurement error and uncertainty, and (c) the role of measuring scales and units in the evaluation of magnitudes.

SUMMARY

We congratulate all of the authors in this volume for the thoughtful, informative, and interesting chapters that they have produced for the fifth edition of *Educational Measurement*. We are hopeful that you, as readers, are both stimulated and informed by their contributions as you

approach the many complex challenges you must meet as you go about applying the science of educational measurement in the service of students, teachers, and the general community, whether as practitioners, educators, policy makers, or journalists.

ADDENDUM

Since the drafting of Chapter 1 of this volume, Donald J. Trump was inaugurated as the 47th President of the United States on January 20, 2025. According to Time Magazine, “the first 100 days of his second term have been among the most destabilizing in American history, a blitz of power grabs, strategic shifts, and direct attacks that have left opponents, global counterparts, and even many supporters stunned” (Cortellessa, 2025, para. 4).

Trump’s actions should not be a surprise to anyone. He has set about methodically addressing his campaign promises as documented in Project 2025, which Wikipedia describes as “a political initiative to reshape the federal government of the United States and consolidate executive power in favor of right-wing policies” (“Project 2025,” 2025, para. 1). One of those promises resulted in the creation of the Department of Government Efficiency (DOGE; The White House, 2025a), which was operational by the end of January 2025. Elon Musk, reputed to be the richest man in the world, was appointed by Donald Trump to lead DOGE, with the goal of cutting federal expenses by at least 30 percent. At the time we are writing this addendum, no federal agency has gone untouched. However, we will focus here only on cuts to agencies and funding that are impacting education and educational research.

We want to be clear that the purpose of this addendum is not to provide an extensive review of all of the funding cuts and threats to educational research and educational institutions brought about by the Trump administration’s policies. Our purpose is to provide a few examples that will help our readers understand the significant impact

the federal government is currently having on society and particularly on educational research.

One of the early cabinet appointments that Donald Trump made was the appointment of Linda McMahon as Secretary of Education (U.S. Department of Education, n.d.). Trump signed an executive order directing McMahon to “take all necessary steps to facilitate the closure of the Department of Education and return authority over education to the States and local communities” (The White House, 2025b, Sec. 2).

On March 11, 2025, McMahon announced a reduction in force of the Department that resulted in nearly 50% of the department’s work force being placed on administrative leave (U.S. Department of Education, 2025). At the time we are writing this addendum, the Institute of Education Sciences (IES), the research arm of the Department of Education, has seen budget and contract cuts that amount to hundreds of millions of dollars. Staff laid off at IES include the teams that administer tests such as the National Assessment of Educational Progress (NAEP), which is considered a gold standard of assessments and is the only national test that compares the performance of students across all 50 states, and the Programme for International Student Assessment (PISA), which measures the performance of U.S. students to allow comparisons to those around the world (Goldstein & Mervosh, 2025, para. 12).

The cuts to educational research supported by the federal government through the Department of Education have resulted in large numbers of researchers no longer able to complete their research or to find work in their field. In addition to cuts in federal funding that deal directly with education, large amounts of funding have been withdrawn from colleges, universities, and institutions carrying out research in a variety of areas. Withdrawal of research funds is causing universities to rescind admissions offers and to freeze hiring. The Trump administration has tied government funding to the need for universities to aggressively curb antisemitism and to be more open to conservative ideas. In March of 2025, the Trump administration cancelled \$400 million in grants and contracts to Columbia University alleging that the school had failed to fight antisemitism (Otterman & Stack, 2025).

On May 20th, 2025, The Harvard Crimson reported that the U.S. Department of Health and Human Services cut “an additional \$60 million in multi-year grants to Harvard over allegations of campus antisemitism and racial discrimination . . . the Trump administration’s third funding cut to Harvard in the last three months” (Patel & Yoon, 2025, para 1). This cut came in addition to the \$450 million cut announced the previous week and the \$2.2 billion cut announced in April by eight federal agencies. The university has sued the Trump administration over two of the funding cuts (Patel & Yoon, 2025).

In addition, very recently, the department of Homeland Security has revoked Harvard’s ability to enroll international students—which currently make up more than 25% of its enrollment—under the F-1 visa program (Diaz, 2025, p. 15). Secretary Kristi Noem said that Harvard was being targeted because they tolerate

“anti-American, pro-terrorist agitators” on campus (Mangan, 2025, para. 2). A temporary restraining order allowing international students to remain enrolled at Harvard pending a court hearing was issued. The university has filed a lawsuit challenging the ban on enrollment of international students (Mangan, 2025).

As well as making cuts in funding to colleges and universities, the Trump administration threatened to “withhold federal funding from public schools unless state education officials verified the elimination of all programs that it said unfairly promoted diversity, equity, and inclusion” (Bender, 2025, para. 1). In the same article, Bender also pointed out that “government-downsizing efforts led by Elon Musk . . . has said that it canceled 70 grants in the Education Department for D.E.I. [diversity, equity, and inclusion] training programs worth \$373 million” (para. 15).

We hope the few examples that we have provided will give our readers a sense of the political climate that is impacting educational researchers and educational institutions in the U.S. today. What we have not been able to describe adequately here is the fear in the educational research community and in U.S. colleges and universities that academic freedom, as we have known it, may be eroding. These fears are related to what is perceived to be punishment of both students and faculty members involved in protests that unfolded in the spring of 2024 over the Israeli-Palestinian conflict. Many protests were construed as antisemitic and have led directly to some of the losses in funding that we have described above. Students and faculty at universities who have participated in these protests have also suffered personal damage (Hartocollis, 2024).

Given the fast pace at which the Trump administration’s actions have unfolded, what we have written here is most assuredly obsolete, and may have been only days after we wrote it. For example, Elon Musk has left his quasi-governmental role, and there is strain between him and the president (Pager & Schleifer, 2025). As mentioned previously, our intent was not to provide an exact history of what has transpired in the first half of 2025, but to give a sense of the rapid changes and the resulting challenges posed in our field. It is unclear what the future of educational research, the institutions that support it, and public education may be in the U.S. We hope, however, that some good will come of the challenges we are witnessing now and that educational research and research institutions will emerge stronger as a result of the many stresses that are being placed on them.

REFERENCES

- American Library Association. (2023, May 9). *Banned and challenged books*. American Library Association. <https://ala.org/bbooks/book-ban-data>
- Anderson, J. (2022, February 23). *The state of critical race theory in education*. Harvard Graduate School of Education. <https://www.gse.harvard.edu/ideas/edcast/22/02/state-critical-race-theory-education>

- Bender, M. C. (2025, April 3). Trump administration threatens to withhold funds from public schools. *The New York Times*. <https://www.nytimes.com/2025/04/03/us/politics/public-school-funding-trump-dei.html>
- Bill of Rights Institute. (2024, May 3). In *Wikipedia*. <https://enWikipedia.org/wiki/Bill-of-Rights-Institute>
- Birnbach, S. (2021, August 15). *Combating racism—Critical race theory*. <https://sarahbirnbach.com/2021/08/15/combating-racism-critical-race-theory/>
- Black Lives Matter: About. (2024). *Black lives matter*. <http://blacklivesmatter.com/about/>
- Branch, G., & Reid, A. (2020, September 12). Evolution education in the U.S. is getting better. *The Scientific American*. <https://www.scientificamerican.com/article/evolution-education-in-the-u-s-is-getting-better/>
- Camara, W. T. (2020). Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practices*, 39(3), 10–18.
- Center for Reinventing Public Education. (2021). *How has the pandemic affected students' social-emotional well-being? A review of the evidence to date*. Mary Lou Fulton Teachers College, Arizona State University.
- Cortellesa, E. (2025, April 25). Exclusive: Inside Trump's first 100 days. *Time Magazine*. <https://time.com/7280106/trump-interview-100-days-2025/>
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.
- Cronbach, L. J. (1989b). Review of *Educational Measurement*, third edition. *Educational Measurement: Issues and Practice*, 8(4), 22–25. <https://doi.org/10.1111/j.1745-3992.1989.tb00339.x>
- Diaz, J. (2025, May 24). By the numbers: A look at international students at Harvard and across the U.S. *NPR*. <https://www.npr.org/2025/05/24/nx-s1-5409082/international-students-numbers-harvard-yale-columbia>
- Edley, C., Jr., Koenig, J., Neilsen, N., & Citro, C. (Eds.). (2019). *Monitoring educational equity*. *National Academies of Sciences, Engineering, and Medicine*. The National Academies Press. <https://doi.org/10.17226/25389>
- Education Recovery Scorecard. (2024, February 28). *New data shows how the pandemic affected learning across whole communities*. Harvard Graduate School of Education. <https://www.gse.harvard.edu/ideas/news/23/05/new-data-show-how-pandemic-affected-learning-across-whole-communities>
- Elias, S., & Feagin, J. R. (2016). *Racial theories in social science: A systemic racism critique*. Routledge.
- Forgash, E. (2024, May 9). Shafik writes op-ed urging universities to do “serious soul searching” about their role in free speech, global conflicts. *The Columbia Spectator*. <https://www.columbiaspectator.com/news/2024/05/09/shafik-writes-op-ed-urging-universities-to-do-serious-soul-searching-about-their-role-in-free-speech-global-conflicts>

- Goldstein, D., & Mervosh, S. (2025, March 13). Are schools succeeding? Trump education department cuts could make it hard to know. *The New York Times*. <https://www.nytimes.com/2025/03/12/us/education-department-cuts.html>
- Hambleton, R. K., & Zenisky, A. I. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. <https://doi.org/10.1037/14049-023>
- Harris, E. A. (2024, April 9). Here are the most targeted books of 2023. *The New York Times*. <https://www.nytimes.com/2024/04/08/books/banned-books-2023.html>
- Hartocollis, A. (2024, October 23). Professors in trouble over protests wonder if academic freedom is dying. *The New York Times*. <https://www.nytimes.com/2024/10/23/us/faculty-protests-academic-freedom-tenure-discipline.html>
- Hartocollis, A., & Fawcett, E. (2023, February 9). The College Board strips down its A.P. curriculum for African American studies. *The New York Times*. <https://www.nytimes.com/2023/02/01/us/college-board-advanced-placement-african-american-studies.html>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Koretz D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education/Praeger.
- Kuhfeld, M., Soland, J., Lewis, K., & Morton, E. (2022, March 3). *The pandemic has had devastating impacts on learning. What will it take to help students catch up?* Brookings. <https://www.brookings.edu/articles/the-pandemic-has-had-devastating-impacts-on-learning-what-will-it-take-to-help-students-catch-up/>
- Levenson, M. L. (2022, May 12). Massachusetts to pay \$56 million after deadly COVID outbreak at veterans' home. *The New York Times*. <https://www.nytimes.com/2022/05/12/us/massachusetts-nursing-home-vets-covid.html>
- List of social movements. (2024, May 3). In *Wikipedia*. <https://en.wikipedia.org/wiki/List-of-social-movements>
- Mangan, D. (2025, May 23). Trump administration ban on Harvard international student enrollment halted by judge. *CNBC*. <https://www.cnbc.com/2025/05/23/harvard-sues-trump-international-student-enrollment.html>
- Masci, D. (2007, June 13). *Twenty years after a landmark supreme court decision, Americans are still fighting about evolution*. Pew Research Center. <https://www.pewresearch.org/religion/2007/06/13/twenty-years-after-a-landmark-supreme-court-decision-americans-are-still-fighting-against-evolution/>

- Mervosh, S., & Goldstein, D. (2023, May 9). Florida rejects dozens of social studies textbooks, and forces changes in others. *The New York Times*. <https://www.nytimes.com/2023/05/09/us/desantis-florida-social-studies-textbooks.html>.
- Mervosh, S., & Wu, A. (2022, October 24). Math scores fell in nearly every state, and reading dipped on national exam. *The New York Times*. <https://www.nytimes.com/2022/10/24/us/math-reading-scores-pandemic.html>
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955–966. <https://doi.org/10.1037/0003-066X.30.10.955>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Murder of George Floyd. (2024, May 7). In *Wikipedia*. https://en.wikipedia.org/wiki/Murder_of_George_Floyd
- Otterman, S., & Stack, L. (2025, March 8). White House cancels \$400 million in grants and contracts to Columbia. *The New York Times*. <https://www.nytimes.com/2025/03/07/nyregion/trump-administration-columbia-grants-cancelled-antisemitism.html>
- Pager, T., & Schleifer, T. (2025, June 6). The unlikely alliance between Trump and Musk breaks down in a rapid and public fashion. *The New York Times*. <https://www.nytimes.com/2025/06/05/us/politics/trump-musk-policy-bill.html>
- Patel, D. T., & Yoon, G. E. (2025, May 20). HHS freezes \$60 million in federal grants to Harvard in third round of Trump cuts. *The Harvard Crimson*. <https://www.thecrimson.com/article/2025/5/20/hhs-60-million-cut/>
- Project 2025. (2025, May 24). In *Wikipedia*. https://en.wikipedia.org/wiki/Project_2025
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.
- Randall, J. (2024, May 9). Jennifer Randall Dunn family endowed professor of psychometrics and test development, Marsal Family School of Education. Marsal Family School of Education. <https://marsal.umich.edu/directory/faculty-staff-jennifer-randall>
- Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, 27(2), 170–178. <https://doi.org/10.1080/10627197.2022.2042682>
- Russell, M. (2024). *Systemic racism and educational measurement*. Routledge.
- Schwartz, S. (2021, June 11). Map: Where critical race theory is under attack. *Education Week*. <https://www.edweek.org/policy-politics/map-where-critical-race-theory-is-under-attack/2021/06>
- Scopes trial. (2024, May 7). In *Britannica Online Encyclopedia*. <https://www.britannica.com/event/Scopes-Trial>
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19(1), 405–450.

- Sireci, S. G. (2020). Standardization and UNDERSTANDARDization in educational assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105.
- The White House. (2025a, January 20). *Establishing and implementing the President's "Department of Government Efficiency"* (Executive Order). <https://www.whitehouse.gov/presidential-actions/2025/01/establishing-and-implementing-the-presidents-department-of-government-efficiency/>
- The White House. (2025b, March 20). *Improving education outcome by empowering parents, states, and communities* (Executive Order). <https://www.whitehouse.gov/presidential-actions/2025/03/improving-education-outcomes-by-empowering-parents-states-and-communities/>
- United Nations. (2006). *Social justice in an open world: The role of the United Nations*. Department of Economic and Social Affairs, Division of Social Policy and Development.
- U.S. Department of Education. (n.d.). *Linda E. McMahon, Secretary of Education*. <https://www.ed.gov/about/ed-organization/meet-secretary-of-education/linda-e-mcmahon>
- U.S. Department of Education. (2015). *Every Student Succeeds Act (ESSA)*. <https://www.ed.gov>essa>
- U.S. Department of Education. (2025, March 11). *U.S. Department of Education initiates reduction in force*. <https://www.ed.gov/about/news/press-release/us-department-of-education-initiates-reduction-force>
- U.S. Department of the Treasury. (2020). *About the CARES Act and the Consolidated Appropriations Act*. <https://home.treasury.gov/policy-issues/coronavirus/about-the-cares-act>
- Vo, T. T. (2021, July 7). A social justice critical framing for opportunity-to-learn indicator use. *The CenterLine Newsletter*. <https://www.nciea.org/blog/a-social-justice-critical-framing-for-opportunity-to-learn-indicator-use/>
- Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton University Press.

NOTES

1. John Pitoniak, the uncle of Mary Pitoniak, was one of at least 84 veterans, most of them elderly, who died at the Holyoke Soldiers' Home after an outbreak of COVID-19 in March 2020. The Commonwealth of Massachusetts acknowledged in the settlement the "deliberate indifference" shown by administrators who combined infected or symptomatic veterans in a unit with those who did not have symptoms (Levenson, 2022).
2. According to Gloria Ladson-Billings, renowned theorist, author, and teacher, "critical race theory is a theoretical tool that began in legal studies, in law schools, in an attempt to explain racial inequity. It serves the same function in education. How do

you explain the inequity of achievement, the racial inequity of achievement in our schools?” (Anderson, 2022, Transcript section, para. 8).

3. #BlackLivesMatter was founded in 2013 in response to the acquittal of Trayvon Martin’s murderer. Black Lives Matter Global Network Foundation, Inc., is a global organization in the United States, the United Kingdom, and Canada whose mission is to eradicate White supremacy and build local power to intervene in violence inflicted on Black communities by the state and vigilantes (*Black Lives Matter: About*, 2024).
4. George Perry Floyd Jr. (October 14, 1973–May 25, 2020) was a Black American man who was murdered by a White police officer in Minneapolis, Minnesota, during an arrest made after a store clerk suspected Floyd may have used a counterfeit \$20 bill (“Murder of George Floyd,” 2024).
5. A movement of mostly young people dedicated to end the era of fossil fuels and support unions and Black, Brown, and working-class communities.
6. A cultural movement that advocates slowing down the pace of human life.
7. The periods are defined by the “growth model pilot program of 2005” (<https://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/index.html>), the “waiver” period beginning in 2011 (Congressional Research Service), and the Every Student Succeeds Act of 2015 (<https://www.everystudentsucceedsact.org>).
8. In Operation Varsity Blues, parents allegedly paid an admissions consultant between \$200,000 and \$6.5 million to guarantee their children’s admission to several different colleges and universities through a variety of means.

The History of Educational Measurement

Brian E. Clauser

NBME (Retired)

Jerome C. Clauser

American Board of Internal Medicine

Amanda L. Clauser

NBME

The history of educational measurement is not a simple history of evolving theory leading to the current state of the art. It is, in fact, multiple separate but related histories. There is a history of test theory and a related history defined by the key figures in a range of fields. Some of these figures, such as Charles Spearman, made critical contributions to test theory; others, such as Walter Bingham and Lewis Terman, influenced testing practice but are likely to go unmentioned in a discussion of test theory. Gustav Fechner studied psychophysics, Alfred Binet studied child development (and mesmerism), and Spearman explored the nature of intelligence; together, they provide a significant part of the foundation on which educational measurement has developed. Other important contributors, such as Truman Kelley, appear to have been singularly focused on the development of methodology.

In addition to these parallel histories built around theory and important contributors to the field, there is a history of evolving technology. Computers have become so widespread in test administration that someone new to the field might overlook the impact that the optical scanner had on testing practice during the second half of the 20th century. In addition to the obvious impact that technology has had on the scoring and administration of tests, computational power has profoundly impacted psychometric theory and practice. In the days when *computers* were individuals using slide rules or mechanical calculators, even a correlation coefficient might have required hours—or days—of work. Pearson (1907), Spearman (1904a), and Kuder and Richardson (1937) all put substantial effort into developing computationally simple approaches to approximating correlation (or reliability) coefficients. The increase in computational power since the 1930s has made such approximations seem quaint and has brought with it applications of Bayesian statistics, hierarchical models, and machine learning that have made educational measurement into something that would be almost unrecognizable to early practitioners.

There has also been an evolution in the application of testing. Spearman was interested in correlations across tests within populations as a means of understanding the nature of intelligence (Spearman, 1904a). Much early large-scale testing—in the military and industry—focused on personnel selection. In this context, tests evaluated individual differences. This same perspective accompanied the introduction of large-scale testing programs in schools, including not only achievement tests, which remain common, but also group IQ tests. Admissions tests follow a similar model; more recently, considerable emphasis has been given to national and international assessments for which individual scores may not even be calculated.

Finally, there is also a history related to the social trends and pressures that have provided the context in which educational measurement has evolved. The Darwinian revolution of the second half of the 19th century led to the eugenics movement, which had widespread support until the Second World War. Much of educational measurement in the early 1900s became inextricably tied to that movement. Similarly, it does not seem like unjustified speculation to suggest that the emphasis that Cronbach (1971, 1989b) and Messick (1975, 1989) placed on fairness and consequences as part of test validity

can be accounted for (at least in part) by the social context created by the civil rights movement and the war on poverty. These social pressures have also included an ebb and flow of the presence of the antitesting movement.

In considering the history of educational measurement, a similar challenge is raised by the question of where to begin. The use of examinations to select individuals for government positions in China began more than 3,000 years ago. In 1115 BCE, candidates were tested in the “six arts”: music, archery, horsemanship, writing, arithmetic, and rites and ceremonies of public and private life. During the Han dynasty—approximately 2,000 years ago—written examinations were introduced in the “five studies”: civil law, military affairs, agriculture, revenue, and geography. After the 7th century, national civil service tests were introduced for use across the empire. These tests emphasized the ability to remember and interpret the Confucian classics. They used both written and oral formats (Zhang & Luo, 2020). Modeled at least in part on these Chinese tests, in 1833 a testing program was introduced for selecting individuals for the British Indian civil service. Testing at universities as part of the process of granting degrees has also been an established practice for centuries (DuBois, 1970).

The breadth and depth of this history makes it clear that a comprehensive history is well beyond the scope of this chapter. Instead, we have identified specific parts of that history which we believe are relevant for understanding the current state of educational measurement. In selecting the parts we chose to stress, we considered several competing priorities. Most people find origin stories interesting. We hope that the events and individuals we discuss create a compelling story that will motivate some readers to explore this history in more detail. In this regard, we hope our chapter will provide a basis for basic historical literacy related to the field of educational measurement. In addition, we have attempted to provide context for understanding some of the major developments in both the theory and the practice of educational measurement. We are more likely to use measurement theories appropriately if we know the context in which they were developed. The reliability coefficient has become a ubiquitous part of test theory and practice, but our reliance on this coefficient may well have as much to do with the history of the field as it does with the usefulness of the coefficient (Cronbach & Shavelson, 2004). Additionally, viewing measurement and testing in the context of an early history that is closely tied to intelligence testing and eugenics is likely to help us understand the skepticism of and resistance to testing that have been present since the 1920s.

It is likely too much to hope that understanding the role that the misuse of measurement has played in supporting injustice will ensure that measurement is not misused in the future, but such an understanding is likely to help us remember that we need to ask not only what we *can* do with our technology, but also what we *should* (and should not) do.

With these considerations in mind, we have divided this chapter into eight sections, each with a different focus. The first section provides a discussion of some of the early foundations of educational measurement, including the contributions of Charles Spearman, Gustav Fechner, and Alfred Binet. To understand how Spearman’s contributions

came into being, we begin with the intellectual culture of the middle of the 19th century that led Francis Galton to study heredity and how Galton's use of correlation in that study provided the statistical tool that Spearman would subsequently apply to test scores. Fechner is important because his work on psychophysics represents what is likely the first scientific effort to measure psychological phenomena. Similarly, Binet introduced not only the intelligence test, but also, along with Theodore Simon, the idea that plots of performance against an independent variable (in this case, age rather than proficiency) can be used in test development.

The second section examines the rise of intelligence testing in the United States. We begin with the work of Henry Goddard and Lewis Terman in the first quarter of the 20th century. Their early work positioned them to be leaders in the development of the army testing program during World War I. That program provided a significant impetus for the rise of testing in the United States. The rise of intelligence testing in turn provided impetus for the eugenics movement in the United States.

The third section provides an overview of the impact that U.S. government programs have had on shaping the practice of educational measurement. This includes the related military research carried out during and after World War I. It also includes legislation that has funded education and both mandated and defined the scope of educational assessment.

The fourth through sixth sections trace the development of three critical aspects of test theory: classical test theory, scaling, and item response theory (IRT). The fourth section follows the development of classical test theory, picking up where the first section left off. This includes the work of Kelley, G. Frederic Kuder, and Marion Richardson, as well as Lee Cronbach's work on coefficient alpha, concluding with a discussion of the development of generalizability theory.

The fifth section builds on the discussion that was begun in the first section, which introduced the development of scaling through the contributions of Fechner and Binet. This section continues the story, following it through to the work of Edward Thorndike and Louis Thurstone, who together laid a kind of theoretical foundation for IRT. The sixth section discusses the independent but parallel development of IRT and Rasch measurement.

The seventh section again shifts from measurement theory to practice, examining the history of the use of large-scale tests both for selection and for accountability. We begin with the tests that Horace Mann introduced to evaluate the performance of the Boston school system in the first half of the 19th century. We then show how this type of test evolved to produce assessments like the College Board examinations and truly large-scale assessments, including the SAT (formerly Scholastic Aptitude Test) and the National Assessment of Educational Progress (NAEP).

The eighth and final section in this chapter shows how the five editions of *Educational Measurement* both document and have been a part of the history of educational measurement.

EARLY FOUNDATIONS OF EDUCATIONAL MEASUREMENT

As we noted in the introduction, educational measurement has a long history. Written standardized tests for selection of government officials were administered in China 3,000 years ago. Written tests have been used at universities for hundreds of years—the competitive mathematics test administered at Cambridge has been in place since the 18th century. Tests for the British Indian civil service began in 1833. These early tests for selection and graduation laid a foundation for the developments that followed. They used standardized conditions of administration and emphasized fairness and accuracy in scoring, but as DuBois (1970) noted, testing during this period lacked the theoretical framework and statistical techniques that have become the hallmark of educational measurement. Those conceptualizations had their roots in England, Germany, and France in work that spanned the period from 1860 through 1910.

England can appropriately be viewed as the birthplace of correlational psychology, and much of what we refer to as classical test theory has its beginning in the work of Charles Spearman (1863–1945), the best known of the early proponents of correlational psychology. Germany was the birthplace of experimental psychology. Gustav Fechner developed psychophysics procedures at the University of Leipzig and, in the process, provided what was likely the first scientific measure of psychological phenomena. Finally, in France, Alfred Binet developed an intelligence test with the practical goal of identifying children in need of special schooling.

Developments in England

In Harold Gulliksen's 1950 text, he summarized the history of classical test theory in a single sentence: "Nearly all the basic formulas that are particularly useful in test theory are found in Spearman's early papers" (p. 1). In a very real sense, Spearman created the foundation for what we know as mathematical test theory, so we begin this history by describing the historical context that brought Spearman's formulas into being.

The chain of events that led to Spearman's work began nearly half a century earlier and halfway around the globe. In 1858, Alfred Russel Wallace (1823–1913) was working in the Malay Archipelago. Wallace supported himself by collecting natural history specimens to be sold in England; he was also collecting evidence in an effort to answer the question of how species come into being. That year, he wrote a paper titled *On the Tendency of Varieties to Depart Indefinitely From the Original Type* (Wallace, 1858). That paper provided an elegant summary of Charles Darwin's—unpublished—theory of natural selection. Wallace mailed his manuscript to Darwin (1809–1882) with the request that if he thought it was of merit, he should forward it to the eminent geologist Charles Lyell. At the point at which Darwin received the letter, he had spent much of the previous 2 decades working (largely in secret) on the same theory and collecting material for a planned multivolume work on the topic.

The events that followed the arrival of Wallace's letter are well known. Darwin shared Wallace's paper with Lyell and Joseph Hooker, two men who were both preeminent members of the British scientific community and Darwin's closest friends. They arranged to have Wallace's paper published with two short pieces by Darwin (Brooks, 1984). All of this has importance to the history of test theory because it made Darwin realize that he could no longer delay publication of a more complete statement of his theory. In just over a year, the first edition of *On the Origin of Species* was published (Darwin, 1859). That book profoundly impacted the history of science; it also profoundly impacted Darwin's first cousin, Francis Galton.

Francis Galton

Francis Galton (1822–1911) was the quintessential Victorian polymath. He studied meteorology, created some of the first weather maps, and discovered the phenomenon of the anticyclone. He investigated the use of fingerprints for identification and wrote three books that were instrumental in the adoption of fingerprint technology (Galton, 1892, 1893, 1895). He explored a part of Africa previously unknown to Europeans (Galton, 1853). He wrote extensively on eugenics (Galton, 1909); in fact, he coined the term. But of all the seemingly limitless areas that Galton explored, his greatest contributions came from his study of heredity motivated by his reading of Darwin's *Origin*.

Galton had begun his intellectual life by studying medicine at King's College London. He had pursued medicine at the urging of his father; he also studied mathematics at Cambridge. After his father's death, Galton used his inherited wealth to mount an expedition to Africa. When he returned, he published a book on his travels (Galton, 1853) and was awarded a medal by the Royal Geographical Society. But as Stephen Stigler (1986) wrote, "Darwin's theories opened an intellectual continent" (p. 267) for Galton to explore far greater than Africa; Galton spent decades mapping that intellectual continent. This exploration of heredity led him to identify the phenomenon of regression to the mean and develop the powerful analytic tool represented by *correlation*.

Galton began his work on inheritance by documenting the extent to which exceptional talent was shared across generations in the same family. In *Hereditary Genius* (1869), he recorded instance after instance in which notable individuals (in the arts, science, and jurisprudence) have notable relatives in the same field. He used this evidence to argue that intellectual and personality characteristics can be inherited in the same way that physical characteristics are passed from parent to child. From a contemporary perspective, the book appears simplistic; Galton too quickly discounted the impact of environmental factors and personal advantage as explanations for his results. Within the context of the times, the work must have been more impressive. Darwin appears to have overlooked these alternative explanations when he wrote to Galton (Darwin, 1869, first paragraph):

I have only read about 50 pages of your Book (to the Judges) but I must exhale myself, else something will go wrong in my inside. I do not think I ever in all my life read anything more interesting & original. And how well & clearly you put every point!

Galton was an astute observer; he also carried out organized experiments to understand inheritance. In 1875, he experimented on pea plants. (Gregor Mendel's work had been published 9 years earlier in an obscure journal [Mendel, 1866], but remained unknown to the broader European scientific community.) Galton described the experiment in his autobiography (Galton, 1908).

The following question had been much in my mind. How is it possible for a population to remain alike in its features, as a whole, during many successive generations, if the average produce of each couple resemble their parents? . . . I was very desirous of ascertaining the facts of the case. After much consideration and many inquiries, I determined . . . on experimenting with sweet peas, which were suggested to me both by Sir Joseph Hooker and by Mr. Darwin. . . . I procured a large number of seeds from the same bin, and selected seven weights. I persuaded friends living in various parts of the country, each to plant a set for me. . . . The result clearly proved *Regression*; the mean Filial deviation was only one-third that of the parental one, and the experiments all concurred. The formula that expresses the descent from one generation of a people to the next, showed that the generations would be identical if this kind of *Regression* was allowed for. (pp. 300–302)

Galton followed the sweet pea experiments with efforts to collect data on humans. He solicited help from friends to collect physical measurements on parents and children and on siblings. He also opened the anthropometric laboratory at the International Health Exhibition in London. This allowed him to collect a range of measures on individuals who visited the exhibition (Galton, 1884).

These data collection efforts allowed Galton to refine his understanding of regression, and in 1886 he contributed two papers on familial relations to the Royal Society (Galton, 1886a, 1886b). These focused his attention on tables representing the relation between deviations in measures such as stature of the adult child and the same measures for parents. Galton realized that a relation existed between the measures, but he was at a loss to quantify it. Again, he described the event in his autobiography: "At length, one morning, while waiting at a roadside station near Ramsgate for a train, and poring over the diagram in my notebook, it struck me that the lines of equal frequency ran in concentric ellipses" (Galton, 1908). He returned to London and visited the Royal Institution in search of information on conic sections. At the Royal Institution, a chance encounter with the physicist James Dewar led to the suggestion that Dewar's brother-in-law, the mathematician J. Hamilton Dickson, might be able to help. Dickson viewed it as a simple problem and his solution was presented as an appendix to Galton's resulting paper.

Two years later, Galton (1888) introduced the correlation coefficient. At that time, the term and general concept of correlation were well established. As Galton stated,

“co-relation or correlation of structure” is a phrase much used in biology, and not least in that branch of it which deals with heredity . . . ; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree. (p. 135)

The 1888 paper built on his earlier work on regression. Again, using anthropological data, he showed how a coefficient could be produced by placing the measures of the correlated attributes on the same scale. This coefficient had the advantage that:

(1) $y = rX$ for all values of y ; (2) that r is the same, whichever of the two values is taken as the subject; (3) that r is always less than 1; (4) that r measures the closeness of co-relation. (p. 145)

Galton had succeeded not only in making an important contribution to the understanding of heredity, but also in providing the mathematical tools to explore a wide range of phenomena. Although we have found no references to indicate that Galton was aware of Spearman's use of the correlation coefficient, if he was aware of Spearman's work, he likely would not have been surprised. In the introductory chapter of *Natural Inheritance*, Galton (1889) commented that it would be worth the reader's time to understand the methodology he used in his study of heredity because of its wide application:

It familiarizes us with the measurement of variability, and with the curious laws of chance that apply to a vast diversity of social subjects. This part of the inquiry may be said to run along a road on a high level, that affords wide views in unexpected directions, and from which easy descents may be made to totally different goals to those we have now to reach. (p. 3)

Galton's great contribution to statistical science—and the history of educational measurement—was the phenomenon of regression to the mean and the powerful analytic tool correlation. There is, of course, some controversy about attributing the discovery of correlation to Galton, but at a minimum, he deserves credit for pointing out the broad usefulness of the procedure in the social and biological sciences. His own words written in another context describe his role in the history of correlation well: “It is a most common experience that what one inventor knew to be original, and believed to be new, had been invented independently by others many times before, but had never become established” (Galton, 1889, p. 33).

In addition to making direct contributions to the development of various fields of science, Galton impacted scientific development through the influence he had on the next generation of thinkers. One individual strongly influenced by Galton was Francis Edgeworth; Edgeworth made important contributions to the development of the correlation coefficient and wrote what may be the first discussion of the impact of measurement error on test scores; we will discuss his work in detail in a subsequent section. Another important follower of Galton was Karl Pearson.

Karl Pearson

Galton's concept of correlation is critical to classical test theory, but the specifics of the coefficient Galton created have long since been replaced by the more mathematically tractable form presented by Karl Pearson. (Galton's approach was based on median and interquartile distance rather than mean and variance.) Although Pearson spent relatively little time studying psychological measurement, his development of the product moment correlation coefficient and his relationships with Galton and Spearman make him a critical part of this history nonetheless.

Like Galton, Pearson (1857–1936) studied mathematics at Cambridge. Unlike Galton, who suffered a breakdown attempting to earn an honors degree, Pearson excelled in those studies, ranking third “wrangler” in the competitive examination taken by all Cambridge undergraduates studying math(s).¹ Pearson was an incredibly productive researcher and author: The annotated bibliography of his works contains 648 entries (Morant & Welch, 1939). Those hundreds of publications included important contributions to statistics; he is responsible for the chi squared test (Pearson, 1900) and for the product moment correlation coefficient (Pearson, 1896). Pearson presented the correlation coefficient in the following form,

$$R = S(xy)/(n\sigma_1\sigma_2) \quad (1)$$

crediting others with developing this equation; he used the bulk of his 1896 paper to argue for the advantages of this formula.²

Pearson devoted much of his professional life to advancing Galton's ideas. With Galton's support, he founded and edited the journal *Biometrika* to advance the statistical study of evolutionary biology. After Galton's death, Pearson held the Galton chair in eugenics, created with an endowment from Galton.

Pearson's admiration for Galton is clear in the following quotation from a speech Pearson delivered at a dinner in his own honor in 1934:

“Road on a high level,” “wide views in unexpected directions,” “easy descents to totally different goals”—here was a field for an adventurous roamer! I felt like a buccaneer of Drake's days. . . . I interpreted that sentence of Galton to mean that there was a category broader than causation, namely correlation, of which causation was only the limit, and that his new conception of correlation brought psychology, anthropology, medicine and sociology in large parts into the field of mathematical treatment. It was Galton who first freed me from the prejudice that sound mathematics could only be applied to natural phenomena under a category of causation. Here for the first time was a possibility—I will not say a certainty of reaching knowledge—as valid as physical knowledge was then thought to be—in the field of living forms and above all in the field of human conduct. (Pearson, 1934, pp. 22–23)

This speech also provides a vivid reminder of another part of Galton's legacy that Pearson embraced:

Buccaneer expeditions into many fields followed; fights took place on many seas, but whether we had right or wrong, whether we lost or won, we did produce some effect. The climax culminated in Galton's preaching eugenics, and his foundation of the Eugenics Professorship. Did I say "culmination"? No, that lies rather in the future, perhaps with Reichskanzler Hitler and his proposals to regenerate the German people. In Germany a vast experiment is in hand, and some of you may live to see its results. (Pearson, 1934, p. 23)³

Correlational Psychology in 1900

Eight years after Pearson published his paper on the product moment correlation coefficient, Spearman began publishing the papers that Gulliksen referred to as containing the formulas that are particularly useful in test theory. The content of those papers warrants special attention, but before examining that content, it is appropriate to consider the state of correlational psychology immediately before the publication of Spearman's papers. Spearman's methodological contributions to educational measurement were enormous, but he was not the first to apply correlational methods to psychology.

The first of these earlier papers we consider was written by William Bagley. Bagley (1901) tested schoolchildren using a range of what might be called physical and mental tests. In general, he found no relationship (or inverse relationships) between these measures. He specifically noted, "There seems to be little direct relation between mental ability as represented by reaction times, and mental ability as represented by class standings" (Bagley, 1901, p. 205). Although Bagley's paper has the title *On the Correlation of Mental and Motor Ability in School Children*, his approach to "correlation" was simplistic even in the context of the times. He examined the relationships between his measures by dividing the first measure into quintiles and averaging the values within each quintile. He then averaged the corresponding values of the second variable for the students in those same quintiles. The five pairs of numbers were then examined directly or plotted to determine whether any relationship existed.

Two additional papers from *Psychological Review* also provide examples of the prior application of correlation in psychology. Clark Wissler's Columbia University dissertation was published in 1901. S. Lovie and Lovie (2010) suggested that this was likely the first use of the product moment correlation coefficient published in a psychology journal. Wissler's paper was an early example of the now-proud tradition of using undergraduates in psychology research. Starting with the hypothesis that different measures of intelligence would be correlated, he reported on numerous measures of physical and mental proficiency, including reaction time, accuracy in marking out As in text, speed in marking out As in text, accuracy in drawing a line, length and breadth of head, and class standing in numerous subjects. Contrary to Wissler's hypothesis, the strength of these relationships was modest except for that between class standing in different subjects. Wissler recognized that "failure to

correlate” may be “due to want of precision in the tests” (Wissler, 1901; p. 29) but gave no serious consideration to how reliability might be calculated or used to adjust the correlation.

The second of the two papers (Aikens et al., 1902) also presents relations between measures of proficiency. The research shows a remarkable lack of statistical sophistication. The “correlations” that are presented have no clear relationship to the Pearson product moment approach. Instead, they used what might be described as a make-shift approximation. Again, the results show generally modest relations between the measured proficiencies, and again, the paper lacks any consideration of how to evaluate the impact of systematic or random effects that might impact the correlation of interest.

Charles Spearman

As a young man, Spearman had an interest in Indian philosophy and joined the army with the hope of being stationed in India. In his 1930 autobiographical chapter, Spearman described the nearly 15 years he spent in the army as wasted, although it did give him the opportunity to complete a 2-year “staff college.” Not long after completing these studies, he resigned his commission and moved to Leipzig to study psychology with Wilhelm Wundt. His studies were interrupted when he was recalled to the army at the outbreak of the Boer War, but eventually he returned to Leipzig, completed his doctorate, and subsequently became a professor at University College London (P. Lovie & Lovie, 1996; S. Lovie & Lovie, 2010; Spearman, 1930).

Spearman’s work focused on understanding the nature of intelligence. Although his most enduring contribution to science may be the methodological tools he developed for correlational psychology and test theory, these tools were developed to support empirical evaluation of the nature of intelligence and specifically to explore the theory that there is a general factor that underlies all measures of intelligence. Differences in reliability across tests impacted the observed correlations between test scores. Much of Spearman’s methodological work was focused on accounting for this effect.

We began this discussion of the developments that took place in England with Gulliksen’s reference to Spearman’s early papers (Gulliksen, 1950). The first of these papers was “The Proof and Measurement of Association Between Two Things” (Spearman, 1904b). This was followed later in the same year by “‘General Intelligence’ Objectively Determined and Measured” (Spearman, 1904a). Both were published while Spearman was still working on his dissertation in Leipzig. Interestingly, these papers were unrelated to his dissertation work, which focused on spatial localization, a topic more in line with Wundt’s program of research (Jensen, 1998; Tinker, 1932). The first paper introduced variations on the rank-order correlation coefficient and argued for the advantages of correlation based on ranks rather than measures. It also presented Yule’s (1897) formula for partial correlation, which Spearman stated he derived using

an independent approach. Most importantly, the paper introduced the formula for estimating true score correlation (p. 98)⁴:

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}. \quad (2)$$

In this equation, r_{pq} represents the true score correlation between variables p and q , $r_{p'q'}$ is the observed correlation between p and q , and the denominator on the right side of the equation represents the square root of the product of the reliabilities for p' and q' .

In discussing the importance of correction that this formula produced, Spearman quoted a paper in which Pearson had concluded that “the mental characteristics in man are inherited in precisely the same manner as the physical” (Spearman, 1904b, pp. 97–98). Spearman pointed out that if the observed correlations between relatives are similar (which Pearson reported), the actual relationship must be substantially stronger for the psychological characteristics because they are measured less precisely. This criticism started a feud between the two University College London professors that continued for nearly a quarter of a century.

Pearson’s immediate response to Spearman’s criticism was to republish the lecture Spearman cited, adding an addendum responding to the criticism (Pearson, 1904). The addendum attacked Spearman’s correction for attenuation because it was capable of producing results that exceed unity and called on Spearman to provide algebraic proofs of the formulas. Spearman responded to Pearson’s criticism in 1907 with a “Demonstration of Formulae for True Measurement of Correlation.” That same year, Pearson published a paper on methods for estimating correlation and again attacked Spearman’s work. In Pearson’s 39-page report (Pearson, 1907), Spearman’s name appears more than 20 times. This creates the impression that the real motivation behind the publication may have had more to do with attacking Spearman than with advancing the practice of correlational science.

The last of Spearman’s papers to make a major contribution to test theory, “Correlation Calculated From Faulty Data,” was published in 1910. It again responded to Pearson’s criticisms and, more importantly, presented the Spearman–Brown formula,

$$r_{x[p],x[p]} = \frac{p \cdot r_{x[q],x[q]}}{q + (p - q)r_{x[q],x[q]}}. \quad (3)$$

In this formulation presented by Spearman (1910, p. 281), $r_{x[q],x[q]}$ represents the reliability of a test with $2qi$ items (with i being any number), and $r_{x[p],x[p]}$ represents the reliability when the test comprises $2pi$ items.⁵ This formula represents a seminal contribution to test theory. When the test length is doubled, the formula provides a basis for estimating split-half reliability. As we will see, this framework led to conceptualizing reliability in terms of interitem correlation and ultimately led to coefficient alpha.

Immediately following Spearman’s paper in the *British Journal of Psychology* was a paper by William Brown (1881–1952; Brown, 1910a), which presents an alternative proof for the same formula. Because of this, the names Spearman and Brown have

been linked as though they were coauthors; that is an ironic misconception. The work presented in Brown's paper comes from his then-recently completed dissertation. The dissertation was also published by the author (Brown, 1910b) and in modified form as *The Essentials of Mental Measurement* (Brown, 1911). Much of the dissertation—and the subsequent book—was an attack on Spearman. Both publications so carefully documented Pearson's criticisms of Spearman that even without Brown's acknowledgment—"Professor Pearson has very kindly read the entire thesis in proof, and made several useful suggestions" (Brown, 1910b, p. 1)—it would be clear that Pearson and not Spearman had influenced Brown's work.

Spearman's 1910 paper is a response to Pearson's criticisms, and he clearly hoped to minimize their differences, concluding, "On the whole, if we eliminate these misapprehensions and oversights, there seems to be no serious difference of opinion on all these points between Pearson and myself" (p. 288). If this was intended as an olive branch, it was one that Pearson did not accept; Pearson continued his attacks for years to come (e.g., Pearson & Moul, 1927).

That said, there is one area in which Pearson and Spearman seem to have been in agreement: They both fully supported the eugenics movement. The following quote concisely captures Spearman's views: "One can conceive the establishment of a minimum index (of general intelligence) to qualify for parliamentary vote, and above all, for the right to have offspring" (Hart & Spearman, 1912, p. 79).

Spearman's Immediate Impact on Measurement

Despite Pearson's criticism, Spearman's ideas became widely accepted, although it is difficult to make a clear statement about how fast that acceptance came about. Edward Thorndike, Wilfrid Lay, and P. R. Dean published a paper in 1909 calling the conclusions of Spearman (1904a) into question, but that same paper applied Spearman's correction for unreliability without comment. The authors appeared to accept Spearman's methodology while rejecting the idea that the available evidence supported the existence of "general intelligence."

Two years later, A. R. Abelson—also at University College London—published an extensive study on the development of a battery of tests. In this paper, Abelson used Spearman's methodology and provided what may be the first example of the use of reliability coefficients to determine the number of items to include on a test form (Abelson, 1911).

By the middle of the 1920s, numerous researchers had published empirical studies evaluating the accuracy of the Spearman–Brown formula. Using a variety of tests, the researchers compared the predicted reliabilities to actual results produced by changing the test length: Holzinger and Clayton (1925) did so using the Otis advanced mental test, Ruch et al. (1926) used spelling words, and B. D. Wood (1926) used achievement tests. The results left little doubt about the usefulness of the formula. Remmers et al. (1927) extended these findings by applying the Spearman–Brown formula to rating

scales and showing that the formula predicted the change in reliability as a function of the number of raters completing the scale.

As we will discuss in a subsequent section, the formulations provided by Spearman did much to shape thinking about reliability for the next 50 years. Numerous researchers made substantial contributions, but the general approach based on correlation coefficients and a conceptualization of reliability as split-half reliability remained intact.⁶

Psychophysics and the University of Leipzig

It would be difficult to write a history of educational measurement and not mention the University of Leipzig. The fact that Charles Spearman received his doctorate from Leipzig would be enough to ensure inclusion, but the university made a foundational contribution to what would become educational measurement well before Spearman arrived. Psychophysics can reasonably be viewed as the starting point for the empirical study of psychology. Gustav Fechner's (1801–1887) 10-year project that culminated in his *Elements of Psychophysics* (1860/1966) may reasonably represent the first effort to measure the psychological. In the process, he laid the foundations for psychological scaling. Stigler (1999) also credited Fechner's work as being the reason that psychology was the first of the social sciences to incorporate statistical theory.

Fechner was born in 1801 in Lower Lusatia, part of the Holy Roman Empire. He moved to Leipzig at the age of 16 to study medicine and stayed for the rest of his life. For the history of educational measurement, and more broadly psychology, the critical events took place in 1850 when Fechner apparently had a kind of epiphany. As Stigler (1986) stated, the stories of that epiphany are “colorful and dramatic” (p. 243). The dramatic aside, Fechner's interest not only in physics but also in philosophy and psychology led him to think about the relationship between physical energy and psychological energy; his insight was that if it is possible to measure physical energy, it must also be possible to quantify the corresponding psychological energy. Fechner took inspiration from the work of Ernst Weber (1795–1878), who was also at the University of Leipzig. Weber had done earlier work on discriminating between physical stimuli and had concluded that the magnitude of a just noticeable difference was a fixed proportion of the base stimulus (Boring, 1950). The study of just noticeable differences can take many forms. For example, a subject might be asked to lift two weights—which outwardly appear identical—and report on whether they are of equal weight or if one is heavier than the other. Obviously, if the difference in the two weights is small enough, it will generally not be perceived. As the difference increases, the probability that it will be perceived increases. Weber's law (sometimes known as the Weber–Fechner law) posits that the ratio of the difference between the two weights (ΔI) and the base weight (I) is a constant (k):

$$\frac{\Delta I}{I} = k. \quad (4)$$

Put another way, as the base weight of the objects increases, the difference in weight between the objects must increase proportionally to have the same probability of being

noticed. As we will discuss in more detail in the section on scaling, Fechner modified this relationship by basing it on the logarithm of the stimulus.

For the bulk of Fechner's work in this area, he experimented with discriminating between weights. He used apparently identical containers, one of which represented the base weight and the other of which would have an additional weight placed inside. For a given pair of weights, the relevant metric was the number of correct responses divided by the number of trials. The resulting Weber-Fechner law was a central result reported in the *Elements of Psychophysics* (Fechner, 1860/1966), but the real contributions of Fechner's work are methodological. Over time, both the accuracy and the generalizability of the law have been called into question, but the general approach Fechner used provided a model for subsequent work in psychophysics, and more importantly, Fechner introduced research design and statistical analysis into the practice of psychology.

Fechner recognized that a wide variety of factors could impact his results: which hand was used, which weight was presented first, the subject's level of fatigue. To account for these sorts of effects, he used balanced factorial designs that allowed him to not only account for the effect across repetitions of the experiment, but also directly estimate the effect of different conditions. For example, he might have a subject lift the same two weights with both the right and the left hands. Although Fechner did not use random sampling, Stigler described the *Elements of Psychophysics* (Fechner, 1860/1966) as the most comprehensive treatment of experimental design to appear before R. A. Fisher's (1935) *Design of Experiments* was published 75 years later.

In imagining Fechner's (1860/1966) experiments, it may be easy for the reader to envision a laboratory in which undergraduates queued up to participate in replications in which they lifted weights. This would be far from the reality. Much of Fechner's work was carried out with Fechner as the only subject—a subject who knew which container contained which weight.

After Fechner retired from lecturing, Wilhelm Wundt came to the University of Leipzig. Wundt continued the psychophysical program of research and established the Institute for Experimental Psychology. This would become the world's first graduate program in psychology.

Alfred Binet

In the early 1900s, when Spearman was publishing the papers that provided the foundation for test theory, Alfred Binet (1857–1911) was publishing a series of papers that provided a foundation for the practice of intelligence testing. Although Binet worked as a psychologist for decades, these papers and the scale they describe are the reason his name is so widely recognized.

Binet studied law; he received his license in 1878, but was unhappy with his career choice and began reading psychology at the Bibliothèque nationale, where he became familiar with Galton's work. In 1882, he met Jean Martin Charcot and Charles Féré and began working in Charcot's laboratory observing and experimenting with the effects of hypnosis on hysteria. In 1887, Binet published a monograph with Féré entitled *Le mag-*

netisme animal. During that year, he also became a student at his father-in-law's laboratory of embryology at the College de France. At the laboratory, he spent several hours per week practicing dissection. The results of this work provided the basis for Binet's doctoral dissertation, *A Contribution to the Study of the Subintestinal Nervous System of Insects* (Wolf, 1973).

During this time, Binet acted as an outspoken supporter of Charcot's ideas about hypnosis. As the debate over these ideas became more heated, the relatively young Binet found himself in an unenviable and untenable position and eventually severed his relationship with Charcot. In search of a new institutional affiliation, Binet requested a staff position at the Laboratory of Physiology and Psychology at the Sorbonne. In 1894, he became director of the laboratory and in collaboration with the previous director founded the first French psychology journal, *L'année psychologique* (Wolf, 1973).

During the latter part of his career, Binet had several less senior collaborators; the most noteworthy of these was Theodore Simon. When Simon approached Binet, he had completed medical school studies and needed a thesis to finalize his degree. At the time, Simon was an intern at the colony for children and adolescents with intellectual disabilities at Perray-Vaucluse. Rather than accept Simon into the laboratory outright, he challenged him to return to the colony and take a series of measurements of each of the 223 boys in the institution and report descriptive statistics for each measure by age group. Several months later, Simon returned, with his analyses completed, and was accepted by Binet; the measurements formed the basis of the thesis for his medical degree.

At the same time that Simon became Binet's assistant, Binet was asked to become part of the *Societe libre pour l'etude psychologique de l'enfant*. The society was subsequently asked to form a commission on the education of children with intellectual disabilities. Part of the commission's charge was to develop a means of identifying children who would benefit from being educated outside the regular classroom. As Binet and Simon described it (1916),

In October 1904 the Minister of Public Instruction named a commission which was charged with the study of measures to be taken for insuring the benefits of instruction to defective children. . . . They decided that no child suspected of retardation should be eliminated from the ordinary school and admitted to a special class, without being subjected to a pedagogical and medical examination from which it could be certified that because of the state of his intelligence, he was unable to profit, in an average measure, from the instruction given in the ordinary schools. (p. 9)

They go on to say that such decisions about placement

have at all times too much the nature of the arbitrary, of caprice, of indifference. Such a condition is quite unfortunate because the interests of the child demand a more careful method. To be a member of a special class can never be a mark of distinction, and such as do not merit it, must be spared the record. (pp. 9–10)

As a leader of the society, Binet was appointed to a commission and spent much of the remainder of his career (and his life) working on this problem.

Between 1905 and 1911, Binet and Simon published a series of papers in *L'année psychologique* that outlined the problem and recorded the effort they made to develop an instrument to identify children with intellectual disabilities.⁷ They began by reviewing the state of the art for such diagnosis, arguing that at the time diagnosis was typically carried out by physicians who based their conclusions on physical symptoms. Binet and Simon asserted that causes such as birth trauma or hydrocephalus may result in disability, but they do not explain the extent of the intellectual deficit. They cited a number of diagnostic frameworks with categories such as *hydrocephalic idiocy*, *epileptic idiocy*, and *traumatic idiocy*. They then argued that even when previous researchers had provided diagnostic criteria based on observable manifestations of intelligence, the criteria had been too vague to be useful. For example, one system stated that “in profound idiocy ‘the attention is fugitive,’ while in imbecility, ‘the attention is fleeting’” (Binet & Simon, 1916; p. 23).

In their next paper, also in 1905 (included in Binet & Simon, 1916), Binet and Simon laid out a synthetic approach including psychological, pedagogical, and medical methods. They began their description of the psychological method by admitting the limitations on measurement when it is applied to intelligence: “The scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not super-posable, and therefore cannot be measured as linear surfaces” (Binet & Simon, 1916; p. 40).

They noted that previous researchers had built their approaches on theory. By contrast, Binet and Simon made practical decisions in constructing their scale. They tried many tests and retained those that proved useful. They designed their individual tests to be administered quickly to avoid fatigue or loss of interest on the part of the child. Finally, they described what we would now refer to as standardized conditions for the tests: “The examination should take place in a quiet room”; “the child should be called in without other children”; when the child is introduced to the experimenter, “he should be reassured by the presence of someone he knows”; and so on (Binet & Simon, 1916; p. 44).

Binet and Simon then presented procedures for carrying out 30 tests. At the simple end of the continuum, the experimenter was instructed to move a lit match across the subject’s field of vision and record whether he or she followed it visually. They assessed “recognition of food” by alternately presenting a piece of chocolate and a similarly shaped piece of white wood. The most sophisticated test required the subject to define abstract terms: for example, “What difference is there between weariness and sadness?”

Binet and Simon followed this description of their procedures with a 90-page report on applications of their new method. Binet’s contribution to educational measurement may be viewed as more of a contribution to practice than to theory. That said, the presentation of the original Binet–Simon scale included methodological and conceptual breakthroughs that have been enormously important. The first of these is empirical eval-

uation of tests (items) for inclusion in the scale. Binet and Simon tried out numerous tests and retained those that discriminated between children of different ages. Although they did not have sophisticated mathematical models to guide them, their approach to evaluating test material was a precursor to the item characteristic curves that would become an essential part of IRT.⁸

The second conceptual contribution that Binet and Simon made was the idea of a norm-referenced interpretation of the scores. Although the samples that were used for the development of their original scale were small—often in the range of 8 to 15 children for a specific age group—they attempted to use children who would be described as average for their age. Admittedly, this seems quaint when compared to the approaches to norming examinations that were used later in the century, but conceptually Binet and Simon provided the starting point for this practice.

The remainder of the work on the Binet–Simon scale was published in 1908 and 1911 (see Binet & Simon, 1916). Both papers describe revisions. In the present context the details are unimportant, but there are two issues that remain salient in the context of intelligence testing. The first is the recognition that these tests are sensitive to education; they are not tests of innate intelligence. One motivation for revision was to remove (sub)tests that were particularly sensitive to education. The second issue is the extent to which scores are influenced by “social conditions.” In the 1911 paper (included in Binet & Simon, 1916), they discussed the work of other researchers who applied their methods and found that many of the children performed well above their age level. They hypothesized that the differences resulted from the fact that the second set of children were from more affluent conditions. Clearly, concern about bias in intelligence testing has a long history.

One final contribution from Binet is an early conceptualization of what we now think of as adaptive testing. Binet recognized that there was little value in administering a test that was very difficult or very easy for the child. He instructed the administrator to move quickly through tasks that were easily mastered by the child or skip them altogether if other evidence from interacting with the child made it clear that the tests would not be useful. Clearly, this practical approach that focused on maintaining the child’s attention is a far cry from the statistically sophisticated approaches that are currently available for adaptive testing, but it shows that an understanding of the value of matching item difficulty to test-taker proficiency was present from the early part of the 20th century.

Taken together, the work that went on over a century ago in England, Germany, and France did much to provide the theoretical foundation of educational measurement. In the next section, we describe how that foundation led to the widespread use of intelligence tests in the United States and how that use has impacted and continues to impact the field of educational measurement.

THE RISE OF INTELLIGENCE TESTING IN AMERICA

Clearly, the development of educational measurement in the 20th century was tightly linked to intelligence testing. Spearman’s main interest in test scores was as a means

of understanding human intelligence, and Binet's motivation was the assessment of intelligence in schoolchildren. In addition to these more obvious connections, one of the earliest efforts to systematically quantify intelligence had its methodological roots in Leipzig.

James McKeen Cattell

In 1882, James McKeen Cattell (1860–1944) won a fellowship to study philosophy at Johns Hopkins. At Hopkins, he worked in G. Stanley Hall's psychology laboratory, where he measured the time in milliseconds it took subjects to recognize letters of the Latin alphabet (Sokal, 1990). After a year at Hopkins, Cattell moved to Leipzig to work on a doctorate at Wundt's laboratory, again studying reaction time. When he left Leipzig, he moved to England to study medicine. In England he met Galton, who was actively working on collection and analysis of anthropometric data (Galton, 1884). Both the reaction time work and anthropometric measurement influenced Cattell's contribution to the history of intelligence testing. In 1891, Cattell moved to Columbia University and established a psychology laboratory; 3 years later, he began testing each student entering Columbia College. The testing program went on into the beginning of the new century. Ultimately, some of these data were analyzed by Clark Wissler (1901) using product moment correlations. The results (which again showed few relations among scores from the different tests) put an end to Cattell's program of research and substantially deterred the use of psychophysical measurements in subsequent intelligence testing (Sokal, 1990). These results ended Cattell's testing program but did nothing to undermine the perceived need for measures of intelligence. As we discussed in the previous section, in 1905 Binet and Simon began publishing work to fill that need, and it was not long before Henry Goddard started to apply Binet and Simon's methods in the United States.

Henry Goddard

Like Cattell, Henry Goddard (1866–1957) studied with G. Stanley Hall, earning his doctorate under Hall in 1899. In 1906, Edward Johnstone, superintendent of the New Jersey School for Feeble-Minded Boys and Girls in Vineland, offered to build Goddard a laboratory if he would come to Vineland to conduct research. Goddard spent the next 2 years working with the diverse population at Vineland and came to conclusions similar to those described by Binet and Simon in their 1905 paper (see Binet & Simon, 1916). Both in France and in the United States the main responsibility for such children fell to physicians, and in both countries the diagnostic classifications and criteria were all but useless.

In 1908, Goddard set out on a tour of Europe in search of more useful approaches, and while in Belgium he came across the intelligence tests developed by Binet and Simon. Goddard's experience at Vineland put him in a position to see the value of Binet's work, and when he returned to New Jersey he translated the tests and began to use them. The resulting scale agreed with the impressions of the Vineland staff

who had direct experience with the children, and Goddard rapidly became a convert. By 1910, Goddard had used the Binet–Simon method extensively. That year, at Goddard’s urging, the Classification Committee of the American Association for the Study of the Feeble-Minded adopted the method. This was a turning point for intelligence testing because it represented formal acceptance of the approach within the medical community (Zenderland, 1990, 1998).

Although Goddard’s immediate challenge at Vineland had to do with decisions around educational placement for children, his interest in *feeble-mindedness* broadened quickly. In 1911, he used the Binet–Simon test in public schools. By 1913, he had tried out the tests at Ellis Island (Zenderland, 1998). At that point, he had already published his eugenically motivated book, *The Kallikak Family* (Goddard, 1912). Kallikak was a pseudonym for a family that Goddard studied; at the time, Deborah Kallikak lived at Vineland. The book was a kind of mirror image of Galton’s *Hereditary Genius*; instead of tracing recurrent instances of genius within the same family, Goddard traced instances of mental deficiency within a single family. Interestingly, although the book highlights the cost to society of such families, Goddard also advocated for the positive results that would be produced by what he viewed as appropriate education.⁹

Goddard was not alone in his interest in the Binet–Simon scale. In 1911, Lewis Terman published an account of his application of the scale with 400 children. And before Terman published the first version of the Stanford–Binet scale there were at least four other English translations and revisions (Hilgard, 1989). In addition to having continued in use through multiple revisions, Terman’s version of the test popularized the use of the intelligence quotient (IQ).¹⁰ This work by Goddard and Terman positioned them both to make contributions when the entrance of the United States into World War I led to a new application for intelligence testing.

Robert Yerkes and Mental Testing in the Army

There can be little doubt that the large-scale use of group intelligence tests during the First World War did much to raise the profile of psychology as a profession and testing as an accepted part of American life. Several authors have called into question the extent to which the test results were used to make important decisions about recruits (e.g., Gould, 1996; Minton, 1990), but there is no question that the army testing program demonstrated the potential for large-scale testing. Although numerous individuals shaped the project, it was led by Robert Yerkes (1876–1956). When Woodrow Wilson went before Congress to ask for a declaration of war, he also asked that universal conscription be instituted to provide the necessary manpower to wage war. Yerkes, the president of the American Psychological Association (APA), saw an opportunity. It is impossible to know to what extent he saw that opportunity as a chance for psychologists to contribute to the war effort and to what extent he may have been motivated by a desire to raise the

profile of psychology as a profession—and, with it, his own standing. Any combination of these motivations may have existed.

Weeks after the declaration of war and before the draft was approved, Yerkes called a meeting of the APA's executive council in Philadelphia. At the meeting, the council approved the formation of an APA Committee on Methods of Psychological Examining of Recruits. The committee met the next month at Vineland (Samelson, 1990; Yoakum & Yerkes, 1920). The group represented a who's who of intelligence testing, including Yerkes, Goddard, Walter Bingham, and Terman. At this time, intelligence tests were typically individually administered, and the initial plans were based on this model. The committee quickly realized that an alternative was needed and within weeks they had developed a prototype for a standardized, group-administered test that could be efficiently administered and scored.

Although the prototype for the army test was put together relatively quickly, it went through various stages of revision and evaluation. After revision, 10 tests remained. These tests were administered to approximately 5,000 men in the U.S. Army and National Guard, as well as samples from "institutions for the feeble-minded," officer training schools, and colleges and universities. The papers were then sent to Columbia University to be evaluated by a group led by E. L. Thorndike "to check their validity, reliability and significance" (Yoakum & Yerkes, 1920, p. 5).

The proposal for widespread testing of recruits with what was termed the Army Alpha was accepted and a school was established for training in military psychology; by the end of the war on November 11, 1918, approximately 120 officers, 300 enlisted men, and 500 clerks were involved in the examination effort. Yoakum and Yerkes (1920) stated (with impressive precision) that 1,726,966 men were examined.

The Army Alpha

The Army Alpha represented a pivotal point in the development of educational measurement in the United States. It represents the beginning of truly large-scale testing and widespread use of intelligence tests. It also provides results that fanned the flames of the eugenics movement and led to the antitesting movement.

As we noted, the committee working on the test quickly came to the realization that a group test was needed. This decision was critically important, but it was not unprecedented. In 1913, William Pyle published *The Examination of School Children*; this was likely the first group intelligence test. It was based on simple constructed response formats (Weinland, 1973).

A second important contribution that impacted the development of Army Alpha was made by Frederick Kelly. In 1915, Kelly published *The Kansas Silent Reading Test*. He was concerned about the unreliability of teachers' marks; he had written his dissertation on the topic. Kelly also hoped to reduce the time and effort required for administration and scoring of tests. Working at the State Normal School at Emporia, Kansas, and then as dean of education at the University of Kansas, he developed an item type that allowed for selection rather than construction of responses. The following two

items qualify as icons in the field of educational measurement in that they are two of the first multiple-choice questions ever published:

Think of the thickness of the peelings of apples and oranges. Put a line around the name of the fruit having the thinner peeling.

apples oranges

Three words are given below. One of them has been left out of this sentence: I cannot _____ the girl who has the flag. Draw a line around the word which is needed in the above sentence.

red see come

Although some parts of the Army Alpha required simple constructed responses—"How many men are 30 men and 7 men?" (Yoakum & Yerkes, 1920, p. 206)—most used the multiple-choice format.

A final contribution that impacted the development of the Army Alpha was made by Arthur Otis (1886–1964). In 1917, Otis was a graduate student at Stanford University, working with Terman. Otis had developed a group intelligence test and Terman brought a copy of the unpublished manuscript to Vineland (Yoakum & Yerkes, 1920). Yerkes's committee, working at Vineland, reviewed hundreds of published tests and ultimately modeled the Army Alpha on Otis's work. Otis interrupted his graduate studies to participate in the development and administration of Army Alpha. He published his version of the test in 1918 (Otis, 1918a, 1918b). After the war, Otis finished his degree and joined the World Book Company,¹¹ which published the Otis Group Intelligence Scale in 1920 and the Otis Self-Administering Tests of Mental Ability in 1922.

Terman, Peacetime Intelligence Testing, and Controversy

Whether the Army Alpha was used for important decisions and improved the efficiency of the war effort may be in question, but there is little doubt that the demonstration of large-scale intelligence testing provided support for the intelligence testing movement in the United States. Application of testing in the army took a back seat until the United States entered the Second World War, but group tests were rapidly adopted by school systems. By 1922, there were more than 40 group intelligence tests being published and in the 1922–1923 school year more than 3 million students were tested (Weinland, 1973). This number continued to rise and large-scale use of intelligence tests continued for decades.

In addition to publishing extensively and developing tests, in 1921 Terman began what is one of the longest-running longitudinal psychological studies in history.¹² Terman identified 1,000 "gifted" children from California with the intention of following them as they grew. The study was intended to provide information about the extent to which childhood IQ scores predicted later success in education and careers. The first results were reported in 1925 (Terman et al., 1925) and the sixth report on the gifted group in later maturity was published in 1995 (Holahan et al., 1995).¹³

As a, if not *the*, leading figure in intelligence testing at the time, Terman was soon at the center of the controversy that arose around testing. In 1922, Walter Lippmann—

author and cofounder of the *New Republic*—wrote a six-part attack on intelligence testing (Lippmann, 1922a, 1922b, 1922c, 1922d, 1922e, 1922f). Much of the content of the articles was specifically targeted at Terman. Lippmann (1922b) began with an effort to ridicule testing by focusing on the apparent absurdity of using the results of the Army Alpha test to conclude that the average American adult had the intelligence of a 14-year-old. Lippmann (1922c) then raised concerns about the potential of the tests to stigmatize children. He also expressed concern that rather than motivating remediation, low scores may provide justification for failing to provide the resources needed for education (Lippmann, 1922d). He moved on to argue against the conclusion that intelligence—as measured by these tests—is inherited and largely unchanged over time (Lippmann, 1922e). Lippmann concluded by suggesting that psychologists should abandon intelligence testing and “save themselves from the reproach of having opened up a new chance for quackery in a field where quacks breed like rabbits” (Lippmann, 1922f, p. 10).

Considering that they were written for the popular press, Lippmann’s articles (1922a, 1922b, 1922c, 1922d, 1922e, 1922f) included a surprising amount of technical detail. He criticized the appropriateness of the norming group used for the Stanford–Binet (Lippmann, 1922a). He argued that the results of the Army Alpha were manipulated by the decisions made about the time limits for the individual tests and that it was impossible to conclude that the intelligence measured by the tests is inherited because so much learning occurs before the age of 4, when these tests were first administered (Lippmann, 1922b).

Terman’s response to Lippmann began mockingly:

After Mr. [William Jennings] Bryan had confounded the evolutionists, and [the flat earth advocate] Voliva the astronomers, it was only fitting that some equally fearless knight should stride forth in righteous wrath and annihilate that other group of pseudo-scientists known as “intelligence testers.” Mr. Walter Lippmann, alone and unaided, has performed just this service. That it took six rambling articles to do the job is unimportant. It is done. The world is deeply in debt to Mr. Lippmann. So are the psychologists, if they only knew it, for henceforth they should know better than to waste their lives monkeying with those silly little “puzzles” or juggling IQ’s and mental ages. (Terman, 1922, p. 116)

In responding to Terman, Lippmann moved from attacking testing to attacking Terman, with lines such as “Mr. Terman’s logical abilities are so primitive that he finds this point impossible to grasp” (Lippmann, 1923, p. 146). Lippmann’s views aside, intelligence testing and the view that intelligence was inherited continued to grow during the 1920s.

Intelligence Testing and Eugenics

Educational measurement had its foundation in the work of Galton and Spearman, both committed eugenicists. During the first 2 decades of the 20th century, the eugenics movement was becoming well established in both England and the United States. Although we are not aware of evidence to suggest that the Army Alpha was motivated by eugenics, after the war, when detailed accounts of the test were published (Yerkes, 1921; Yoakum & Yerkes, 1920), the results certainly fanned the flames of that move-

ment. Two types of results produced particular controversy. First, the very large sample of recruits allowed for examining differences across race and ethnicity. Substantial differences were apparent between Black and White recruits. Similar differences were also observed between recruits who had recently arrived in the United States from eastern and western Europe. These results could not have been more in line with the expectations of the eugenicists who believed that White western Europeans were of superior stock. Both Henry Goddard and Carl Brigham (1890–1943)—psychologists who had worked on the Army Alpha during the war—republished these results in the form of arguments for a eugenic solution. Brigham’s (1923) *A Study of American Intelligence* was particularly strident.¹⁴

Goddard and Brigham were far from the only psychologists of this era who embraced eugenics. Truman Kelley was a member of the advisory council to the Eugenics Committee of the United States of America (I. Fisher, 1923). Psychologists testified before Congress in support of restrictive immigration laws (although there is little evidence that their testimony was a significant factor in passing these already popular measures; Sokal, 1990). Yerkes and Thorndike were members of a Eugenics Record Office committee on the inheritance of mental traits (Zenderland, 1998). In 1927, eugenics had become sufficiently well accepted that the Supreme Court ruled in favor of a Virginia law allowing for compulsory sterilization of individuals with intellectual disabilities. In writing the nearly unanimous decision, Oliver Wendell Holmes stated that “three generations of imbeciles are enough.”

There can be little doubt that the decision to uphold compulsory sterilization and the restrictive immigration policies did irreparable harm and that intelligence testing and testers were closely aligned with these efforts. Nonetheless, it should be kept in mind that, at the time, these were popular positions. In addition to Oliver Wendell Holmes, the list of public figures who wrote in support of eugenics includes Winston Churchill, Margaret Sanger, and Helen Keller¹⁵—to name just a few. Views about eugenics changed dramatically after the Second World War when the Nazi plan to implement eugenics through mass extermination became public. Although intelligence testing continued, the link to heredity (and eugenics) received much less attention. It is difficult to know what the key figures in testing believed about eugenics in the second half of the 20th century; few advocated for it publicly.

Intelligence and Race

The controversy over intelligence testing came back with force in the late 1960s when Arthur Jensen began publishing on racial differences in intelligence. In 1969, *Harvard Educational Review* published a lengthy article by Jensen along with a series of invited responses from other psychologists. Jensen argued that the failure of compensatory education programs (such as Head Start) to produce lasting effects on IQ and achievement raised questions about the assumptions that supported development of these programs. In subsequent books and papers, Jensen further developed the idea of general intelligence (a view similar to Spearman’s), argued

that a substantial proportion of the variance in IQ is the result of inheritance, and stated that the observed differences in IQ scores across racial groups cannot be explained by test bias.

Jensen's work attracted considerable attention both within the scientific community and with the general public, and he was accused of being a racist. Public speaking engagements were disrupted or canceled because of threats of disruption. At one point his university hired a bodyguard for him and his mail was routinely screened for bombs (Robinson & Wainer, 2006). Over the next 25 years, numerous books were published on both sides of the controversy.

The controversy over IQ inheritance, and differences in racial and socioeconomic groups continued for over 2 decades. Jensen was not alone in advancing what was considered inflammatory research on IQ and race. He was also not alone in publishing outside academic journals in newspapers and magazines that were certain to produce a public response. Richard Herrnstein (1971) published in the *Atlantic Monthly* and Linda Gottfredson (1994) published in the *Wall Street Journal*. As with Jensen, publishing on this topic led to calls for removal from their respective universities. Gottfredson (Wainer & Robinson, 2009) lost funding for her position at Johns Hopkins shortly after she began this line of research. Gottfredson (Wainer & Robinson, 2009) and Jensen (1980) also clearly believed that they were excluded from some of the top-tier journals in their fields, not because of the quality of their research, but because of the controversial nature of their results. This may have motivated their decision to make the discussion not only public but also high profile by publishing outside the academic literature. Alternatively, it may be as Cronbach (1975) wrote:

Scholars typically welcome the widest possible attention to their views because they cherish the ideas and because they prize visibility and influence. What wonder, then, that a scholar given his once-in-a-lifetime moment in the public eye will seek to make the moment memorable? (p. 12)

Although publication in this area has continued (e.g., Gottfredson, 2018), the last high-profile contribution to the public controversy came when Herrnstein and Charles Murray (1994) published *The Bell Curve: Intelligence and Class Structure in American Life*. This book became a *New York Times* bestseller and drew immediate attack within the mainstream press (Gould, 1996). Although this resurgence of the controversy no doubt sold many books and magazines, it is unlikely that it changed many minds. It does seem likely that the controversy created an unpleasant association between testing and racism and may have done more than a little to motivate the current antitest movement.

In this section we saw how the development of the Army Alpha impacted the practice of assessment. In the next section we look more broadly at how other federal programs and legislation have impacted educational measurement.

THE INFLUENCE AND IMPACT OF U.S. GOVERNMENT INVOLVEMENT IN ASSESSMENT

As we discussed in the previous section, World War I provided a context for the introduction of large-scale standardized testing in the United States. The Army Alpha was, however, only one of the government-initiated programs that have impacted the theory and practice of testing. In this section we consider how both ongoing Department of Defense funding and federal legislation have continued to impact educational measurement.

It is not surprising that with the U.S. entrance into the Second World War the army would again turn to psychologists to help with the challenge of selection and placement created by the need to turn millions of enlisted men into a functioning army. By 1941, the idea of using tests for selection and placement in industry (and the military) was well established. There was an opportunity for psychologists to contribute to the war effort and at the same time advance the field. One of the most noteworthy individuals driving that effort was John Flanagan.

John Flanagan and the Aviation Psychology Program

Flanagan (1906–1996) had an impressive academic pedigree. As a high school science teacher, he attended summer psychology courses taught by E. L. Thorndike. He then left his teaching position in the late 1920s to study psychology at Harvard University, earning his degree under Truman Kelley. Flanagan then joined the Cooperative Test Service of the American Council on Education. During his time with the Cooperative Test Service, he produced one of the first significant expositions on test equating (Flanagan, 1939), which developed into the chapter on equating in the first edition of *Educational Measurement* (Flanagan, 1951). This chapter may well have been the first comprehensive presentation of the subject. Flanagan examined the importance of form development and comparable application of test specifications, as well as appropriate assignment of items to forms to support score comparability. He considered statistical approaches to equating as well as providing recommendations for the statistical machinery of equipercentile equating and arguing against the use of regression functions for equating because they provide an asymmetric link between test forms (Flanagan, 1951). Despite these significant contributions to the field, Flanagan is likely best known for his work on psychological testing during World War II, including development of the critical incident technique, which has been widely used in developing specifications for credentialing examinations and is still used in the early 21st century.

Like many of the other figures we have discussed, Flanagan's early work was tied to the eugenics movement. His initial experience in military research began in 1938 when he was hired by Frederick Osborn to conduct a study of the factors responsible for decisions regarding family size made by Army Air Corps officers and their wives (Flanagan,

1942). The foundation commissioning the study was concerned that Army Air Corps pilots—“considered the ‘cream of the crop’ genetically”—were having small families. The intention of the research was to identify “incentives to encourage them to have more children” (Capshew, 1999, p. 107).

This work led to Flanagan’s commission in the U.S. Army Air Corps in 1941 to lead the Aviation Psychology Program. Under his leadership, more than 150 psychologists were similarly commissioned and contributed to the program, with more than 1,400 individuals contributing throughout the war (“APF Gold Medal Awards,” 1993). The extensive work of the Aviation Psychology Program of the Army Air Forces is documented in 19 volumes describing studies in development and analysis of criterion measures, rating scales, checklists, and materials developed based on critical incident reports. The Aviation Psychology Program introduced criterion-referenced testing and test validation approaches based on performance. The validity research included comparisons of test results with data collected during training and subsequent performance in the field (Shields et al., 2001).

After working on the Osborn study, Flanagan was sponsored by the National Research Council to develop the Aviation Cadet Qualifying Examination, a test for classifying aircrew candidates. The examination was initially implemented shortly after the 1941 attack on Pearl Harbor (Staff, Psychological Branch, Office Air Surgeon, 1944). These tests were initially used for cadet selection and subsequently for placement into specific aircrew roles. The novel contribution introduced by Flanagan’s team relied on the rapid development and validation of a complicated assessment battery. The assessment sorted participants into role-specific stanines for navigator, bombardier, and pilot roles, based on a combination of performance-based assessments (e.g., finger dexterity and rudder control), personality, and knowledge/proficiency tests. Validity evidence for the predictive power of these assessments was gathered through training measures (Did the individual successfully complete training?) and measures of success or failure in combat, including ratings, awards, errors, and survival (Flanagan, 1948).

A second research project, conducted during this period under the auspices of the Army Air Forces Aviation Psychology Program, represents a rarely possible textbook example of validity research for a selection procedure. This study was designed to evaluate the procedures and measures for selecting pilots and aircrew to maximize the match between their strengths and their assigned roles. The research reports of the Army Air Forces Aviation Psychology Program (“The Experimental Study of a Thousand Applicants Sent Into Pilot Training”) described how approximately 1,050 men were assigned to pilot preflight school without consideration of their performance on the Army Air Forces qualifying examination battery. They were then enrolled in basic training and primary/advanced flying schools, with some eventually receiving ratings as qualified pilots. “Two-thirds of the men with stanine scores of 8 or 9 completed the program, while fewer than 3% of those with stanines 1 or 2 did so” (Jones, 2007, p. 604). Jones

also noted that the predictive value of the test was substantially reduced when applied to subsequent performance. The more than 20:1 ratio for completing training was reduced to less than 2:1 in a follow-up study predicting accidents associated with pilot error. And although follow-up during combat was plagued with problems arising from incomplete data, it was impossible to demonstrate that higher scores were associated with superior outcomes in combat.

Postwar Developments

The practical work that went on during the war continued to have a significant impact after the war ended. The formalization of the critical incident technique, an evidence-based test design approach used during the war to identify effective actions taken by officers, led to the development of the “critical requirements” of combat leadership for the U.S. Army Air Forces (Flanagan, 1954) and then to a set of procedures that have been broadly used in the construction of credentialing examinations. Robert Thorndike (1947, 1949) went on to devise methodologies for developing a personnel testing program based on his work in the Army Air Forces. These reports detail methods and considerations for performing job analysis, developing aptitude tests (pilot testing, validation, revision), criterion validation, developing composite scores from test batteries, and assessing reliability. Many of the individuals who participated in the Army Air Forces became leaders in the civilian research community—some, like Flanagan and Thorndike, were officers, whereas others, like William Angoff, participated as enlisted men (Jones, 2007).

Department of Defense Support for the Development of Modern Test Theory

Beginning during the Second World War and continuing through the 1990s, the Department of Defense supported a series of projects that advanced the development of what would become known as *item response theory*. The book *Measurement and Prediction* (Stouffer et al., 1950)—published after the war—includes a chapter on Paul Lazarsfeld’s (1901–1976) wartime work on latent structure analysis. That work provided a basis for the conceptual and statistical development of IRT (e.g., Bock, 1997; Lord, 1952). In the 1950s, the U.S. Air Force School of Aviation Medicine supported Allan Birnbaum’s development of the logistic form of the one-, two-, and three-parameter IRT models (Birnbaum, 1957, 1958a, 1958b). The Office of Naval Research supported the writing of Lord and Novick’s (1968) volume that became a nearly sacred text for measurement experts and made Birnbaum’s work widely available. The Office of Naval Research (along with other government agencies) also helped to move IRT from the theoretical to the practical by sponsoring a series of landmark conferences on latent trait theory and computerized adaptive testing. The conferences took place in 1975 (Clark, 1976), 1977 (Weiss, 1978), and 1979 (Weiss, 1983). These conferences represented the state of the art in what became known as IRT and the presenters represented a who’s who of researchers working in that area. The authors listed in these proceedings include Fumiko

Samejima, Frederic Lord, Bert Green, Darrell Bock, Dan Eignor, Ron Hambleton, Mark Reckase, H. Swaminathan, Linda Cook, David Thissen, and Howard Wainer. The Department of Defense also sponsored a more than 30-year-long research project that resulted in computerization of the Armed Services Vocational Aptitude Battery (ASVAB) in 1996. The foreword to a report on this effort described this accomplishment (Sands et al., 1999).

In October 1996, the Department of Defense (DoD) implemented a computerized adaptive testing (CAT) version of its enlistment test battery (the Armed Services Vocational Aptitude Battery or ASVAB) in 65 Military Entrance Processing Stations (MEPSs) across the country. DoD became the first organization to use CAT-derived scores for personnel selection when the system was placed in five MEPSs for operational testing in 1992; now DoD has become the first employer to adopt CAT for its employment system. . . . The Department is the largest single employer of American youth, testing over 350,000 applicants for entrance into the Military Services between October 1, 1994 and September 30, 1995. (p. ix)

The ASVAB project began in the 1960s. Early aspects of this work focused on statistical models for item selection, but subsequent work covered every aspect of implementation, including pool construction, exposure control, software and hardware selection, and validity studies. This effort produced important advances in computerized adaptive testing, but more generally it advanced the field with respect to both technical and practical knowledge of IRT. As important as these contributions were in the evolution of educational measurement, they may well be judged as having a minor impact when compared to the other way the federal government influenced the measurement field—through legislation to support education and assessment.

The Impact of Legislation on Educational Measurement

In response to the Soviet launch of *Sputnik* in 1957 and the perception that America was losing the space race, federal policy began to explicitly support testing of schoolchildren in science, mathematics, reading, and foreign languages through the 1958 National Defense Education Act (NDEA). This served the purposes of establishing baseline knowledge in these areas as well as providing support for rapid curriculum development and implementation in public schools (Bunch, 2021). The testing programs supported by the NDEA were designed to evaluate college readiness, specifically preparedness for postsecondary study in math, science, and foreign languages. The act also provided federal funding for these tests in the event that a state could not legally pay for them (Section 504(b)).

Approaching federal involvement in education as a matter of national defense distinguished the NDEA from previous proposed education-spending legislation, versions of which had been passed by the Senate but languished in the House. Framing support for standard educational goals (and assessments) across states as a matter of national defense was ultimately a winning strategy, although there was controversy about the

appropriate limits and uses of standardized testing in schools. Advocates for this legislation saw the potential to provide greater access to college through federally supported scholarships, loans, and selection instruments, while challengers were skeptical about the utility of standardized tests in making decisions about individual fitness for secondary education or a career path (Urban, 2010).

The programs underwritten by the NDEA continued through the 1960s and eventually were subsumed by the Elementary and Secondary Education Act (ESEA) in 1965. The ESEA was a part of Lyndon Johnson's Great Society initiative and laid the groundwork for state and school district accountability for educating disadvantaged children. The ESEA was reauthorized and/or amended regularly through 2015 (taking on names including No Child Left Behind and Every Student Succeeds Act). The ESEA also brought the federal government into the education of bilingual students and students with disabilities, provided federal money for testing and supplemental educational services for historically disadvantaged students, and established Head Start as a permanent program (Bunch, 2021; Urban, 2010).

A significant part of the motivation to increase the assessment component of this legislation was to ensure that states and localities would effectively implement the requirements for the education of disadvantaged students and students with disabilities. The assessment component was again strengthened with the passage of the 1974 amendments (Education Amendments, 1974) requiring the commissioner of education to develop program evaluation models that "specify objective criteria" and provide "comparable [data] on a statewide and nationwide basis" (Section 151; Bunch, 2021). During the 1970s and 1980s, states were supported by 10 regional technical assistance centers to develop and implement assessments in support of ESEA requirements. These centers employed hundreds of testing specialists and interacted with education officials in all 50 states to create a standardized approach to solving educational measurement problems. Between 1979 and 1981, these centers provided over 5,000 workshops and on-site consultations for more than 80,000 clients (Bunch, 2021; Stonehill & Anderson, 1982).

In addition to promoting assessment through these regional centers, the 1978 reauthorization directed the commissioner of education to assist state education agencies in developing capacity to conduct large-scale achievement testing programs to measure reading, writing, and mathematical proficiency, along with other subjects. After the election of Ronald Reagan in 1980, the ESEA was reauthorized as the Educational Consolidation and Improvement Act. This led to more local control, but somewhat surprisingly it again increased the requirements for assessment. As a result of this act, schools, districts, and states were required to administer criterion-referenced tests to demonstrate that targeted outcomes were being met.

Despite the years of legislation and associated funding, the 1983 report on the state of American education, *A Nation at Risk* (National Commission on Excellence in Education, 1983), described America's schools as failing to educate the nation's children. This led to the 1994 Improving America's Schools Act. The act led to additional assessment requirements: Assessments aligned with the content standards were to be administered

at a point between Grades 3 and 5, between Grades 6 and 9, and again between Grades 10 and 12 (Section 1111(b)(3)(D)). Standards, however, were still established at the state level and were highly variable across states.

In 2002, this trend of increasing emphasis on assessment continued with the passage of No Child Left Behind (NCLB), in place from 2002 to 2015. The act expanded testing to include Grades 3 through 8 and required testing in one grade of high school. As with every reauthorization of the ESEA since 1966, NCLB emphasized the importance of education and testing of minority populations, English language learners, and students with disabilities. NCLB created an emphasis on state standards and assessment that led to meaningful developments in federal education policy. Building on the goal of all children achieving proficiency in reading and mathematics (by the 2013–2014 academic year), NCLB introduced test-based accountability provisions such as annual improvement targets and penalties if targets were not met. NCLB accountability requirements increased policy attention on test accommodations and accessibility, as well as expectations for progress reporting for students and student subgroups.

There can be little doubt that federal legislation has done much to drive assessment of specific populations, influence test accommodations and tested content, and impact how testing is performed in education for students in kindergarten through Grade 12. The legislation described has also shaped the significant role that measurement experts have had in education and accountability efforts since the 1960s. As significant as this has been, it may be that even more psychometric innovation was driven by a different federal initiative, NAEP. We would be remiss not to mention NAEP in this section on federal involvement in assessment, but we will present the details of this program in the subsequent section on large-scale testing.

RELIABILITY THEORY AFTER SPEARMAN

In the prior two sections we focused on the historical and social developments that shaped educational measurement. In this section we return to the more theoretical aspects of the history of educational measurement and follow the development of reliability theory after Spearman's early papers. This development has two parts that are sometimes viewed as chronologically distinct but are, in fact, interwoven. The first has to do with the continued development of classical test theory; the second part resulted in generalizability theory.

The Development of Classical Test Theory

Although Spearman's work provided the foundation for classical test theory, his focus was on creating tools for correlational psychology. He was interested in efficient approaches to estimating correlations between tests and in tools for evaluating those relations. There is little evidence that his interest extended to advancing methodology

for estimating standard errors or otherwise interpreting individual test scores. This was consistent with his interest in using test scores to study intelligence rather than to evaluate individuals. Spearman was a correlational psychologist, not a psychometrician. In this respect, Spearman had a great deal in common with Galton and Pearson. All three were interested in developing correlational tools to advance their respective substantive fields: heredity, evolutionary biology, and human intelligence. The next major contributor to the development of test theory, Truman Kelley, approached testing from a different perspective: that of a methodologist. In fact, Cronbach referred to Kelley as an “obsessive algebraist and statistician . . . who had no motive save methodologic” (Lee Cronbach, personal communication, December 22, 2000).

Truman Kelley

Truman Kelley (1884–1961) earned his doctorate under E. L. Thorndike at Columbia and went on to teach at Stanford and Harvard. In 1922–1923, Kelley spent a sabbatical year with Pearson at the Galton Biometric Laboratory. Around this same time, Kelley began publishing a series of papers that would extend classical test theory beyond the formulas presented by Spearman. What may be Kelley’s single greatest contribution to classical test theory was among his first, the regression estimate of a test taker’s true score:

$$\hat{T}_x = \rho X_i + (1 - \rho)\bar{X}. \quad (5)$$

In this equation taken from Kelley (1923), the left-hand term represents the estimated true score, ρ represents the reliability of the test score, X_i is the observed score for test taker i , and $\bar{X}$ is the population mean. This equation not only represents an important use of the implicit assumptions of classical test theory, but also links test theory to a Bayesian framework by combining the observed data with a prior assumption (Levy & Mislevy, 2021).

The next year, he published a paper showing that if we have two forms of a test such that $x_1 = a + e_1$ and $x_2 = a + e_2$, where x is the observed score, a is the true score, and e is a “chance factor,”¹⁶

$$r_{12} = \frac{\sigma_a^2}{\sigma^2}. \quad (6)$$

In Kelley (1925) he showed the relationship between the reliability of a measure and the correlation between the measure and the related true score:

$$r_{1\infty} = \sqrt{r_{1I}} \quad (7)$$

or, equivalently,

$$r_{1I} = r_{1\infty}^2. \quad (8)$$

These equations will be familiar to anyone who has studied classical test theory; they represent only a sample of Kelley's contribution. In addition to numerous additional papers of this sort, Kelley published what may well be the first book devoted to the mathematical theory of test scores, *Interpretation of Educational Measurements* (1927), and substantially influenced formal presentations of test theory that followed. The extent to which Thurstone's *The Reliability and Validity of Tests* (1931) and Gulliksen's *Theory of Mental Tests* (1950) depend on Kelley's work makes it clear just how essential Kelley's contributions were to the development of educational measurement.

G. F. Kuder and M. W. Richardson

G. F. Kuder (1903–2000) and M. W. Richardson's (1896–1965) 1937 paper followed in the spirit of Kelley in showing how simple assumptions and algebraic manipulation can be used to derive practically useful results. Their paper began with a theoretical discussion of reliability and then progressed through a series of manipulations to produce what have become known as the KR20 and KR21 (the 20th and 21st equations presented in the paper). The first of these represents a generalized equation for reliability and the latter a form that the authors stated could be calculated in 2 minutes using statistics that are readily available to the test developer. Their presentation began with a rejection of the test–retest paradigm for reliability in favor of the split-half approach (highlighting the importance of the Spearman–Brown formula). They noted that split-half methods depend on the specific split used and then presented a general equation that represents the correlation between two equivalent forms of a test. If $a, b, \dots, n$ and $A, B, \dots, N$ are corresponding items in two hypothetical forms of a test, the tests are equivalent if the items in each pair (a and A , b and B , etc.) are interchangeable. Based on this assumption, they present the KR3:

$$r_{tt} = \frac{\sigma_t^2 - \sum_1^n pq + \sum_1^n r_{ii}pq}{\sigma_t^2}. \quad (9)$$

In this equation, r_{tt} is the reliability of interest, σ_t^2 is the variance of the test, p is the difficulty for a specific item, and q equals $1 - p$. They recognized that this formulation was not of practical use because r_{ii} cannot be calculated. After 16 intervening equations, they presented the KR20, derived from KR3 with the assumption that all inter-correlations are equal:

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}. \quad (10)$$

Finally, by assuming that all item difficulties are equal, they produced the KR21,

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}. \quad (11)$$

Kuder and Richardson's (1937) paper was important both practically and theoretically. For practitioners, it provided an approach to estimating the reliability of a test that was computationally simple (at the cost of making strong assumptions). In the days before computers and electronic calculators, this was of considerable value. Their paper also represented an important theoretical step forward in the understanding of reliability.

Lee Cronbach

A final critical step in the development of reliability within the framework of classical test theory was provided by Cronbach's introduction of coefficient alpha. Cronbach (1916–2001) made major contributions to assessment. In addition to his work on advancing the conceptualization of reliability theory, he published groundbreaking work on validity theory. He coauthored the seminal paper on construct validity with Paul Meehl (Cronbach & Meehl, 1955) and wrote the chapter on validity for the 1971 edition of *Educational Measurement*. As an interesting aside, Cronbach was also a member of Terman's gifted group. The gifted group was followed long after Terman's death, and in addition to being a subject in the study, Cronbach assisted with production of the final volume of the study (Holahan et al., 1995).

Cronbach's 1951 paper on alpha may be the most cited paper in the history of measurement. According to Google Scholar, it has been cited tens of thousands of times. In a sense, the coefficient alpha paper extended the presentation from Kuder and Richardson (1937). KR20 is a special case of coefficient alpha in that it applies to items scored 0/1. Cronbach showed that coefficient alpha (and, by extension, KR20) represents the mean of all possible split-half coefficients.¹⁷ It is also "the value expected when two *random* samples of items from a pool like those in the given test are correlated" (Cronbach, 1951; p. 331).

In a posthumously published paper, Cronbach reflected on coefficient alpha (Cronbach & Shavelson, 2004). He argued that although reliability coefficients were appropriate for the kind of correlational psychology that was prevalent from Spearman's time through the first half of the 20th century, more recent applications of measurement are better served with an estimate of the standard error of measurement.¹⁸ In that same paper, Cronbach advocated for procedures based on analysis using variance components rather than correlations. He was arguing that generalizability theory should supplant classical test theory as the preferred framework for conceptualizing reliability.

Generalizability Theory

Cronbach's advocacy for the framework represented by generalizability theory comes as no surprise; he spent over a decade working through the problem of reliability before he and his colleagues published *The Dependability of Behavioral Measurements* (Cronbach et al., 1963, 1972). The work created what the authors referred to as a liberalization of classical test theory.

In classical test theory, observed scores are conceptualized as comprising a true score and an undifferentiated error term. The true score is the expected score that the test taker would receive if they could complete the test an unlimited number of

times. In generalizability theory, the sampling process is conceptualized as occurring at the task (or item) level, with tasks being randomly sampled from a universe of admissible tasks. This allows for multiple facets comprising the total score variance to contribute either to the test taker's universe score (similar to the true score) or to error. The framework allows for conceptualizing the reliability of scores in which (a) separate components of the error variance are associated with the sampling of the items and the raters who evaluate the responses to those items, (b) the total score is a composite based on stratified sampling of items from different content areas or item formats, (c) the measurement is focused on scores for individual test takers or aggregates such as classrooms, and (d) relative versus absolute score interpretations are possible.¹⁹

Clearly, any history of generalizability theory will need to include the work of Cronbach and his collaborators beginning in the late 1950s and lasting very nearly until Cronbach's death, but many of the ideas reflected in generalizability theory first appeared in the late 1800s, and as with classical test theory, they were developed by numerous individuals over time. The earliest of those ideas came from the work of Francis Edgeworth (1845–1926).

Edgeworth attended Trinity College, Dublin, and Oxford University, where he studied ancient and modern languages. He was self-taught in the areas in which he made his reputation: statistics and economics. His work in economics earned him a chair at Oxford. Edgeworth was interested in bringing mathematical analysis into the social sciences. This interest led him to become one of Galton's earliest disciples (Kendall, 1968; Stigler, 1986). This interest also led him to write three papers on the statistics of examinations. Edgeworth's interest expressed in these papers was to examine "the degree of accuracy or inaccuracy which is to be ascribed to the modern method of estimating proficiency by means of numerical marks" (Edgeworth, 1888, p. 600); in other words, he was interested in the reliability of test scores.

When Edgeworth referred to the "modern method," he was speaking of a system in which essays or short-answer papers were scored by content experts. The purpose of the paper was to describe an approach for estimating the extent to which chance influenced examination scores. He began by describing something closely related to what we now think of as the true score in classical test theory (or the universe score in generalizability theory). Using Latin translation as an example, he wrote,

This central figure which is . . . assigned by the greatest number of equally competent judges, is to be regarded as the true value of the Latin prose; just as the true weight of a body is determined by taking the mean of several discrepant measurements. There is indeed this difference between the two species of measurement, that in the case of material weight we can verify the operation. We can appeal from kitchen scales to an atomic balance, and show that the mean of a great number of rough operations tends to coincide with the value determined by the more accurate method. But in intellectual ponderation we have no atomic

balance. We must be content without verification to take the mean of a number of rough measurements as the true value. (Edgeworth, 1888, p. 601)

He went on to define error in terms of variability about this true value and concluded that for aggregate scores error “will fluctuate according to the normal law” (p. 604). Having described the observed score as being composed of a true score and a normally distributed error, Edgeworth then presented a number of the essentials of generalizability theory:

1. He described the total error score as being made up of multiple facets, including what in generalizability theory would be designated as task effects, rater effects, and rater-by-task interactions.
2. He made the assumption that tests were constructed through random sampling from a defined domain—the central assumption of generalizability theory (Kane, 2002)—and commented on conditions that would violate the model.
3. He described the total error as the square root of the sum of the squared effects and noted that the impact of any effect will be reduced by the square root of the number of observations contributing to that effect.

Edgeworth used these conceptualizations to describe data collection designs that would minimize the error without increasing the total amount of rater time required.

Having considered sources of error associated with raters, Edgeworth (1888) then examined the contribution of task sampling to the error of measurement:

We have so far been endeavouring to estimate the error which is incurred in appreciating the actual work done by the candidate. We have now to evaluate the error which is committed in taking his answers as representative of his proficiency. (p. 614)

The suggestion that Edgeworth presented the basics of generalizability theory in this paper may seem like an exaggeration because generalizability theory uses the framework of analysis of variance and R. A. Fisher did not publish his ideas until decades after Edgeworth’s paper (e.g., R. A. Fisher, 1925). It is, nonetheless, obvious that this is the framework Edgeworth intended; he had already published a paper on “analysis of variance” in 1885. Stigler (1986) commented on the 1885 paper that “the solution was insightful and foreshadowed much of twentieth-century work on the analysis of variance” (p. 314).²⁰

Edgeworth’s insights into the generalizability of test scores were impressive. Whether they influenced subsequent work in this area is difficult to say. It may be that, like his contribution on analysis of variance, these papers were largely ignored—perhaps he was too far ahead of his time. We have found no references to his work in the early literature on test theory, but this is questionable evidence because extensive citation of previous work was rare at that time. For most of the half century that followed Edgeworth’s 1888 paper, the emphasis was on the development of classical test theory, which has

already been discussed. The next important contributions to reliability theory that went beyond the classical test theory paradigm were provided in 1936 by Harold Gulliksen (1903–1996) and Cyril Burt (1883–1972).

Gulliksen's (1936) contribution appeared in the first volume of *Psychometrika*. He looked at the typical approach to reliability for objectively scored items; because variability in this context is substantially associated with item sampling, he referred to it as content reliability. He then raised the question of how content reliability can be estimated for essay examinations in which the total error includes both content error and error introduced by "readers." Assuming that two separate forms of the test are available and that one of those forms is scored by two separate readers, he developed an approach to quantifying the content reliability associated with essay tasks.

Burt's (1936) contribution is framed in the context of factor analysis. In his publication, he conceptualized total score as being constructed of multiple factors, including a general factor and sources specific to raters. Using factor analytic and correlational procedures, he described how rater error can be quantified.

Cronbach et al. (1972) described the contributions by Gulliksen (1936) and Burt (1936) as being among the first to conceptualize measurement error as comprising multiple facets. But as Cronbach (1991) would later write, univariate generalizability theory "interweaves ideas from at least two dozen authors" (p. 394). These contributions include R. A. Fisher's work on analysis of variance (R. A. Fisher, 1925), as well as that of Cornfield and Tukey (1956). Burt (1955) also made a second contribution that explicitly linked estimation of reliability to analysis of variance.

These publications and numerous others set the stage for the work Cronbach and his collaborators did in the late 1950s and 1960s. In 1957, Cronbach and Gleser began work on a handbook of measurement theory. Believing that reliability had been thoroughly studied by others, they decided to begin with that section of the book. Cronbach (1991) wrote later, "We learned humility the hard way—the enterprise never got past that topic" (p. 392).

The essential features of univariate generalizability theory were largely completed by the early 1960s and published in three journal articles, each with a different first author (Cronbach et al., 1963; Gleser et al., 1965; and Rajaratnam et al., 1965). These papers incorporated analysis of variance procedures to evaluate multiple facets representing different sources of error, absolute versus relative error, and an understanding of fixed versus random facets in the design. Additional years of effort were then necessary to expand the theory to a multivariate framework. The result was *The Dependability of Behavioral Measurements* (Cronbach et al., 1972).

Their monograph represents a remarkable step forward in both the conceptualization and the estimation of measurement error. Unfortunately, although the volume captured much of the authors' best thinking on the topic, as Cronbach (1991) wrote, "the book proved indigestible" (p. 395). Cronbach et al. (1972) warned that "the reader is certain to gain far more from his third reading of most sections than from his first or second" (p. 4). Although Cronbach continued to work in this area until shortly before his death,

he published relatively little on the topic after the 1972 volume. The extent to which generalizability theory has become widespread seems to a substantial degree to be the result of subsequent authors who have developed the theory and written texts (and software) that make it more accessible. Although numerous researchers have written on the topic, far and away the greatest contribution has come from Robert Brennan. For many years, *Elements of Generalizability Theory* (Brennan, 1983) was the primary introduction that researchers had to generalizability theory. This monograph provided the basics needed for univariate generalizability analysis and prepared readers to make use of *The Dependability of Behavioral Measurements* (Cronbach et al., 1972). Brennan's more complete text and the associated software (Brennan, 2001a, 2001b) provided a detailed resource for multivariate generalizability analysis.

In concluding, we wish to make clear that we do not intend to imply that the technology for evaluating reliability and estimating measurement error is now complete. Instead, our goal has been to explain how this technology has developed during the past century. This summary is intended to describe how these concepts have evolved. Hopefully, this will lead to a better understanding of the current state of the art and will at the same time point to areas for development of both theory and practice.

SCALING

Next, we consider the development of applications of scaling for educational measurement. Unlike the other sections of this chapter, for which the content is likely to be self-evident given the section heading, the term scaling may warrant a definition, or at least a description. This is not because it is an obscure term, but because a focus on scaling has been in and out of vogue over the decades, and ideas about the requirements for a defensible scale have varied as well. In 1954, Lord wrote that “‘scaling’ appears to be virtually indistinguishable in meaning from ‘measurement,’ which may be defined as the assignment of numerals to objects (including people), according to some rule, in order to represent their properties” (p. 375). Lord went on to state, “If a method of scaling rests upon a verifiable hypothesis as to the existence of some specified self-consistency in the data, then the data may or may not be found to scale” (p. 375). Lord's first statement sets a low bar for scaling and, by extension, for measurement. The second implies that if the user establishes a higher bar by way of the intended interpretation of the scores, it may be necessary to demonstrate that the data—and, by extension, the phenomenon represented by those data—support that interpretation. Lord's statement suggests that scaling has occurred even when it is strictly categorical; scales need not be continuous, nor do they need to represent ratio or even interval measurements.²¹

The first two editions of *Educational Measurement* included chapters on the nature of measurement. Consistent with much theory and practice at the time, this chapter did not appear in the third and fourth editions. More recently, the limited attention that has been given to the theoretical requirements of measurement (and scaling) has become a source of some controversy. Michell (1999) has written extensively and crit-

ically on this topic. McGrane and Maul (2020) described the failure to demonstrate that the latent traits reflected in IRT models have characteristics that support the use of the associated scales as a “scientifically dire situation” (p. 8). Recently, several authors have attempted to resolve this controversy by placing scaling procedures in perspective. Briggs (2021) provided a historical perspective incorporating the contributions of Fechner, Galton, Thurstone, and Stanley Stevens. Mari et al. (2021), in their volume *Measurement Across the Sciences*, tried to define the requirements of measurement in psychology from a broader scientific perspective and specifically by bringing in the perspectives of metrology. David Torres Irribarra’s (2021) volume *A Pragmatic Perspective of Measurement* provides yet another view of the controversy bringing both philosophical pragmatism and common-sense pragmatism to the discussion.

For this brief summary of the history of scaling, we focus on developments in practice. We recognize that separating developments in the taxonomy of scaling procedures (see Stevens, 1946, 1951; Torgerson, 1958) and the conceptualization of the nature of measurement more broadly from developments in the practice of scaling is somewhat artificial, but these theoretical issues are presented in detail in Briggs et al. (this volume). To outline the development of scaling, we return to Galton and Fechner, but this time we focus on their efforts to build measures. We then examine how these early efforts at scale development impacted the development of the IRT models that have become ubiquitous. (See Moses, this volume, for a presentation of current approaches to scaling.)

Ernst Weber and Gustav Fechner

In 1860, Gustav Fechner provided an initial formalization of psychological scaling with the publication of *Elements of Psychophysics* (1860/1966). As we described previously, Fechner’s magnum opus was the culmination of a decade of work conducting experiments on individuals’ responses to physical stimuli and extending the work of Ernst Weber. Weber had spent his career examining psychological responses to physical stimuli and had found that when asked to detect differences in physical stimuli, subjects were sensitive to relative changes in proportion, not absolute changes in magnitude. Fechner generalized this work by recognizing that the sensory intensity of a stimulus is the product of the log of the level of physical stimulus multiplied by some constant K (Sensation = $K \log$ Stimulus).²² This finding, later described as Fechner’s law, indicates that there is a logarithmic relationship between a physical stimulus and the associated perception. That is, if you need to double the stimulus to increase the perceived difference from 1 to 2, you must double it again to increase the perceived difference from 2 to 3 (Briggs, 2021).

Although Weber had primarily concerned himself with comparing the magnitude of these proportions across different sensory stimuli, Fechner recognized that these psychological attributes could be expressed on their own scale separate from the scale of the stimulus itself. Fechner noticed in his experiments that subjects sometimes disagreed as to which stimulus was the strongest. He hypothesized that these disagreements were attributable to random error in the assessment of the stimuli. He further hypothesized that these errors in measurement were normally distributed with a mean of zero.

Therefore, subjects disagreed as to which stimuli were more intense because the psychological comparison on the stimuli follow two random variables, each following a normal distribution (Fechner, 1860/1966).

In his experiments, Fechner compared two physical stimuli across many subjects and calculated the proportion of time that the treatment stimulus was judged to be greater than the control stimulus. Then, assuming that the standard deviations of these responses were the same for all pairs and stimuli, Fechner set these equal to 1 and used the inverse of the cumulative density function to compute an estimate of the mean difference in psychological response for those stimuli. After repeating this across multiple pairings of stimuli, Fechner was able to establish a scale for the measurement of sensory intensity.

Fechner's work was important in shaping thinking about the nature of scaling and measurement in psychology—and, by extension, educational measurement—because it established an early precedent that such measurement should follow the framework used in the physical sciences. This goes well beyond Lord's "assignment of numerals to objects . . . according to some rule" (Lord, 1954, p. 375).

Francis Galton

Although Fechner's work is notable as the first attempt to scale psychological stimuli, his work focused on the scaling of stimuli, not subjects. The first significant advance in scaling subjects was achieved through the work of Francis Galton. Although, as we have discussed, Galton is better recognized for his research on inheritance and his discovery of correlation, his work in scaling unobservable human characteristics provided an important foundation for the scaling of mental tests that is still in use in the early 21st century.

Inspired by the work of Adolphe Quetelet, Galton had become fascinated with the normal distribution and its ubiquity in nature (Briggs, 2021; Gillham, 2001). Galton recognized that the distribution of some physical traits, for example, height, was the product of many small, random events in the inheritances of each individual. He believed that unobservable mental traits were equally a product of this random path of inheritances and reasoned that they too should follow a normal distribution. This revelation regarding the shape of the distribution allowed Galton to translate the measure of any individual into its percentile—a word Galton coined—allowing individuals to be placed relative to one another along a statistical scale (Galton, 1875).

Galton referred to this approach to scaling as "relative" and recognized that although it allowed human qualities to be quantified, it lacked an objective reference point. In an effort to ameliorate this limitation, Galton developed scale anchor descriptors to indicate the performance that would be expected at particular percentiles. Furthermore, Galton recognized that relative assessments could be every bit as useful as absolute measurement in the myriad situations where individuals are assessed relative to a peer group, saying that "a blurred vision would be above all price to an individual man in a nation of blind men, though it would hardly enable him to earn his bread elsewhere"

(Galton (1889), p. 36). This relative approach to scaling individuals was valuable for the study of inheritance; by the turn of the 20th century, it was about to make its way into educational measurement. Again, we see physical measurement as the model for psychological measurement.

E. L. Thorndike

The next important contribution to scaling—and the first that was consciously intended to advance the field of educational measurement—was made by E. L. Thorndike (1874–1949). In the early 20th century, secondary education was becoming widespread in the United States. Schools all over the nation needed teachers, and increasingly, large urban universities were adding departments of education to address this need. Columbia University was at the center of this trend, and in 1899 Thorndike became an assistant professor at Columbia Teachers College (Cherry, 2023). Thorndike believed that science was the only sure foundation for social progress and was unhappy with the level of scientific rigor applied to education at that time (Clifford, n.d.). He admired the work of Galton and Pearson and wished to apply the scientific method to the study of education. To this end, he introduced the first course in educational measurement in 1902 and published a textbook, *Educational Psychology*, in 1913.

During his 40 years at Columbia, Thorndike worked on the design and development of many exams for a wide array of purposes (including his evaluation of the Army Alpha described previously). Trained in psychology, Thorndike had limited formal training in mathematics and statistics. Nonetheless, he built on the work of Galton when developing a scaling technique that placed students across grades on a common scale. To accomplish this, Thorndike began by calculating the proportion of test takers answering each question correctly for each grade. Items were then expressed as a deviation from the grade mean in terms of standard deviation units. The mean difference between common items administered in adjacent grade levels was then calculated and used to adjust grade means to place all grades on a common scale (Thorndike, 1919). Although this represented a significant advancement in educational scaling in 1905, this method assumed that the distribution of abilities across grades differed by position but not by dispersion (Engelhard, 1984; Thurstone, 1927). Twenty-five years later, this limitation would be addressed by Thurstone.

L. L. Thurstone

The year that Gustav Fechner passed away in Leipzig, L. L. Thurstone (1887–1955) was born in Chicago, Illinois. While at Cornell studying engineering, Thurstone's interest in education and psychology grew, along with his frustration with the poor quality of teaching in the college of engineering. In 1914, Thurstone abandoned engineering and matriculated at the University of Chicago to pursue his doctorate in psychology (D. A. Wood, 1962). During his graduate training, he accepted an assistantship at the newly created Division of Applied Psychology at the Carnegie Institute of Technology. During the next 9 years at Carnegie, Thurstone was at the forefront of the development

of psychological tests on a wide variety of cognitive functions. In 1924, he returned to the psychology department at the University of Chicago, where he taught a course on test theory.

At this point in his career, Thurstone began to scrutinize the scaling approaches that were common in mental assessments at the time, like those used by Thorndike and others. Thurstone recognized that earlier approaches to educational scaling assumed that dispersion was constant across grade. Thurstone believed this assumption was dubious and in 1924 he articulated a method of scaling that accounts for differences in both group means and dispersion (Thurstone, 1927). Thurstone's absolute scaling approach begins similarly to Thorndike's approach. For each test question, the difficulty is calculated for students in that grade and those difficulty values are rescaled to be expressed as deviations from the group mean in standard deviation units. Unlike Thorndike, however, in the absolute scaling approach, Thurstone calculated that standard deviation of the difficulty for common items in adjacent grades. The ratio of these standard deviations could then be used to rescale one grade onto the scale of the other grade while accounting for the differences in dispersion. This process could be repeated across adjacent grades to create a single scale that spanned several grade levels.

In 1925, Thurstone extended this scaling work to consider placing test takers along these educational scales. To unify the scales for test takers and items, Thurstone placed each test question on the test-taker ability distribution so that the percentage of test takers answering the question correctly was equal to the percentage of test takers to the right of that point in the distribution. This allowed item difficulties to be expressed as the place along the test takers' ability distribution at which 50% of test takers would answer the item correctly. He then represented these item difficulties in terms of standard deviation units, such that items at the mean ability were scaled to zero while more difficult and easier items were expressed as positive and negative normal deviates, respectively. Thurstone went on to graph the empirical probability of a correct response across grade for items at different difficulties and showed that the difficulty distributions are S-shaped, with slopes that vary by item. He stipulated,

A refined statistical method for ascertaining the age at par [item difficulty] for each test question would be to fit an equation to the curve for each test question and then to ascertain the point at which the curve intersects the 50% level. (Thurstone, 1925, p. 446)

Thurstone did not develop that refined statistical method. As we will see in the next section, that development would result in the creation of IRT.

The specific examples of scaling that we provided in this section represent major steps in the development of scaling procedures in educational measurement. Each of these procedures goes well beyond Lord's simple assignment of numerals based on a rule; each assumes "some specified self-consistency in the data" (Lord, 1954, p. 375) that contributes

to the interpretation of the measurements. As the works cited earlier in this section make clear (Briggs, 2022; Irribarra, 2021; Mari et al., 2021; McGrane & Maul, 2020), the defensibility of these types of interpretations remains controversial. The historical perspective provided in this section does not resolve that controversy, but it does provide a context for understanding it. The practical application of these efforts to construct self-consistent scales resulted in IRT—which we consider in the next section—but as we discuss in that section, it did not eliminate the controversy over the requirements for scaling.

ITEM RESPONSE THEORY

The history of IRT is complex. Part of that complexity comes from the fact that IRT as we know it in the early 21st century developed independently—at about the same time—in both the United States and Denmark. In Denmark, that work was the effort of Georg Rasch. In the United States, numerous researchers made individual contributions that supported the ultimate development of IRT.

The Foundations of IRT

Development of the One-, Two-, and Three-Parameter Logistic Models

The lineage of IRT connects it both to the work on classical test theory described earlier in this chapter and to the work on scaling described in the previous section. In 1934, the United States was in the grips of the Great Depression, and the Roosevelt administration was committed to putting Americans back to work through programs like the Works Progress Administration. During this time, Thurstone was teaching psychology at the University of Chicago when the government offered him a new research assistant, Ledyard Tucker (1910–2004; Carlson & von Davier, 2017). Like Thurstone, Tucker had an undergraduate degree in electrical engineering; Thurstone quickly recognized Tucker's talents and encouraged him to continue his education in psychology. After more than a decade collaborating with Thurstone, Tucker earned his doctorate in 1946 and took a position at the College Entrance Examination Board (College Board; Dorans, 2004).

During his brief time at the College Board, Tucker defined the mathematical form for what we now refer to as an item characteristic curve (ICC). Unlike Thurstone, who had calculated nonparametric ICCs using observed data, Tucker used the integral of the normal curve—the normal ogive—to relate test-taker ability to performance (Tucker, 1946). This work made it possible to calculate the probability that a test taker at any ability level would answer a particular item correctly.

Although Tucker described the mathematical form for the ICC, he did not devote much attention to the underlying latent trait, instead relying on the language of true score theory to describe test-taker ability. This next critical step was based on the work of Paul Lazarsfeld (1901–1976), who in his 1950 publication laid the conceptual foundation for the *latent trait*, in what psychometricians soon would call latent trait theory.

Lazarsfeld studied mathematics in Vienna. With the rise of Hitler, he left Austria for the United States and eventually moved to Columbia University, where he became the associate director of the university's Bureau of Applied Social Research ("Dr. Paul Lazarsfeld Dies," 1976). Although Lazarsfeld is known in sociology for his work on voting and the media, his contributions to test theory stem from his development of latent structure analysis. Latent structure analysis posits that there is an unobserved, *latent* cause for a subject's observed responses. By assuming that responses to individual items are independent, Lazarsfeld was able to study the joint occurrences of item response patterns to make inferences about the trait that underlies this behavior (Lazarsfeld, 1950). Although, the majority of Lazarsfeld's work concerned attitude measures, not educational assessments, the reconceptualization of observed data as being the product of a latent trait as opposed to the trait itself was a critical development.

Bert Green (1928–2019) recognized the applicability of Lazarsfeld's latent structure analysis to educational measurement. In the early 1950s, Green set to work analyzing test-taker response data using latent structure and latent class models. He extended Lazarsfeld's work by developing a general procedure for latent structure and latent class models (Green, 1951a, 1951b) and demonstrating the interrelationship between factor analysis and latent structure analysis. Through this work, Green directly connected latent structure analysis models to what would become IRT. The final steps in the introduction of a statistically rigorous theory of item response patterns was then provided by Lord (1952) and Birnbaum (1957, 1958a, 1958b).

In 1944, Frederic Lord (1912–2000) joined the graduate record office of the Carnegie Foundation, the forerunner to Educational Testing Service (now known as ETS). In 1949, he became the director of statistical analysis at the newly formed Educational Testing Service (Carlson & von Davier, 2017). He completed his dissertation in March 1951, and in 1952 the work was published as a psychometric monograph. The monograph presented the one- and two-parameter IRT models based on the normal ogive. Nearly 30 years later he published the first volume that focused on IRT: *Applications of Item Response Theory to Practical Testing Problems* (Lord, 1980). With these two publications—and the numerous journal papers published between them—he may have done more than anyone else to advance the development of IRT.

Lord was a pragmatist who allowed the data to dictate the development of the model. In this vein, Lord was a developer and advocate of the two- and three-parameter IRT models.²³ In his work in the early 1950s, Lord coined several key terms, including ICC and test characteristic curve. He also described many foundational concepts of IRT, including local independence, item invariance, and standard errors conditional on test-taker proficiency (Carlson & von Davier, 2017; Lord, 1952). Lord continued to develop these concepts throughout his career.

Lord deserves considerable credit for development of the two- and three-parameter IRT models based on normal ogive, but it was Allan Birnbaum (1923–1976) who

reconceived Lord's work using the more mathematically tractable logistic models that are common in the early 21st century. In the late 1950s, Birnbaum was introduced to educational measurement when Columbia Teachers College invited him to help support a contract with the U.S. Air Force School of Aviation Medicine on test-taker classification (Barnard & Godambe, 1982). As we have already noted, in earlier work by Tucker, Green, and Lord, ICCs were described using a normal ogive model. Inspired by the work of David Haley (1952) and Joseph Berkson (van der Linden & Hambleton, 1997), Birnbaum believed the logistic form was more appropriate for IRT because of the existence of a sufficient statistic (Barnard & Godambe, 1982). To this end, in the late 1950s Birnbaum published several research reports for the U.S. Air Force on the applicability of the logistic model to IRT. Lord later invited Birnbaum to summarize this work in four chapters in *Statistical Theories of Mental Test Scores* (Birnbaum, 1968). In those chapters, Birnbaum provided the now-familiar logistic forms of the one-, two-, and three-parameter IRT models and introduced the scaling constant D to make the logistic form nearly identical to the earlier normal ogive formulation. Birnbaum also produced a practical method for item parameter estimation and later introduced the use of prior distributions into ability estimation (Birnbaum, 1967).

Birnbaum's contributions can reasonably be viewed as completing the foundations of IRT. As we have mentioned, related work also went on in Denmark. Whereas the models described in the previous paragraphs built on the insights of numerous researchers, the work in Denmark was carried out almost entirely by one man, Georg Rasch.

Development of the Rasch Model

Rasch (1901–1980) was born in Svendborg, Denmark. He studied mathematics at the University of Copenhagen and in 1930 earned his doctorate with the intention of pursuing a teaching position at the university. When Rasch struggled to obtain a professorship in mathematics (during the worldwide depression), he turned his attention to statistical analysis and was able to earn a scholarship to study with the Nobel Prize–winning econometrician Ragnar Frisch and later with R. A. Fisher. He returned to Denmark in 1936 and was invited to teach statistics in the psychology department at the University of Copenhagen (Anderson & Olsen, 2001). It was during this time that he became increasingly interested in the problems of educational scaling. During the next 2 decades, Rasch worked as a statistical consultant, frequently on psychological and educational assessments. In 1952, while analyzing data for a Danish military exam, Rasch began work on the model for which he has become known.

Unlike Lord, who was a pragmatist, Rasch came to believe that it was critical to define the requirements necessary for objective measurement. By the time he published his monograph *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1960/1980), he had developed the principles that have become associated with Rasch measurement: the requirements of sufficient statistics and specific objectivity. It is noteworthy that Rasch's *Probabilistic*

Models did not explicitly include the model widely associated with Rasch today—that is, the model equivalent to the one-parameter IRT model. The equivalent model presented in that volume lacked the logistic transformation. He began by defining the probability of a correct response as the ratio of the test takers' proficiency and the item difficulty.

If we put $\xi/\delta = \zeta$, the problem is to choose a function of ζ which only takes values between 0 and 1. And as we find both very easy and very difficult problems and both very able and very unable persons, ζ must cover all values from zero to infinity.

The simplest function I know of, which increases from 0 to 1 as ζ goes from 0 to ∞ , is $\frac{\zeta}{1 + \zeta}$. If we insert $\zeta = \frac{\xi}{\delta}$ we get $\frac{\frac{\xi}{\delta}}{1 + \frac{\xi}{\delta}} = \frac{\xi}{\xi + \delta}$. (Rasch, 1960/1980; p. 74–75)

If we reparameterize by changing ξ to e^{θ} and δ to e^b , then the model becomes the typical Rasch model or the one-parameter logistic IRT model.

Several authors noted this relationship at approximately the same time. Birnbaum (1968, p. 402) noted that Rasch's model had this characteristic in the context of his more general presentation of the one-, two-, and three-parameter logistic models. Similarly, Wright (1968) included a footnote commenting that Rasch's model could be reformulated in logistic form.²⁴ It is unclear which of these authors has priority in identifying this relationship.²⁵

After Rasch's (1960/1980) monograph, he went on to publish several additional papers on the model. Rasch's Berkeley Symposium paper from 1961 provides a general case for *specific objectivity*, and Rasch (1966) discussed other issues, including conditional independence. Throughout his work, Rasch discussed models that included logistic transformations, so it is a curiosity that he never introduced the now-common formulation of the logistic Rasch model for dichotomous items, $P(\theta) = (1 + e^{\theta - b})^{-1}$. Mark Reckase (personal communication, June 2, 2022) noted that it could be argued that Equation 4.2 in Rasch (1961) is the general case of all the models currently referred to as Rasch models and the simple logistic model is a special case of that equation. It may be that Rasch simply believed that it was more natural to use proficiency and difficulty scales with a lower bound of zero.

Although this reparameterized Rasch model for dichotomous items is equivalent to the one-parameter logistic IRT model, separate terms are often used to describe them to highlight the philosophical distinction between Rasch's approach and that represented by Lord's work.²⁶ Given the attention that has been placed on these philosophical distinctions, the reader might get the impression that Rasch developed his model based on principled a priori assumptions about the requirements for measurement; Rasch's own description of the development of his model does not support that interpretation. Regarding his early work on "reading capability," Rasch stated, "When I got the data I made the guess that it might be a good idea to try the multiplicative Poisson model. It turned out to fit the data quite well" (1960/1980, p. xiii). The con-

cept of *specific objectivity*, which has become a hallmark of Rasch measurement, seems to have been discovered after the fact, rather than motivating the development of the model. Later in life, Rasch made it clear that he only appreciated this characteristic of the model after discussing his work with his former mentor, Ragnar Frisch. Rasch visited Frisch in 1959 and described the work he had been doing in educational measurement. On seeing Rasch's formulation, Frisch expressed particular interest in the idea that the person and item parameters could be separated. This separability is what Rasch came to refer to as specific objectivity. Rasch went on to say, "What Frisch's astonishment had done was to point out to me that the possibility of separating two sets of parameters must be a fundamental property of a very important class of models" (1960/1980, p. xviii).

Rasch Versus Multiparameter IRT

This philosophical fault line that separated IRT and Rasch measurement became a growing fissure in the field of educational measurement from the late 1960s into the 1990s, often producing heated interactions at professional conferences. One such interaction was the 1992 debate between Benjamin Wright (1926–2015) and Ronald Hambleton (1943–2022) at the American Educational Research Association's annual meeting in San Francisco. To provide a sense of the division that existed in the field, we describe that debate in some detail.

Wright and Hambleton met in a packed ballroom to participate in an invited session titled "Which Models Work Best." Wright opened the debate: "I make measures for a living. Measures have a specific definition." He went on to suggest—wrongly—that the Rasch model was developed nearly a decade before multiparameter IRT. He proceeded to criticize multiparameter IRT for its "promiscuity" and willingness to "swallow whatever junk happens to come [its] way." Throughout his remarks, he repeatedly returned to the metaphor of a ruler that fails to provide objective measurement, saying, "no scientist, engineer, businessman or cook, who depends on measures . . . can work with that kind of ruler." He closed his remarks by attacking multiparameter IRT for being data centered, saying that the "Rasch model is theory-centered: data must fit, else get better data."

The younger Hambleton took a more genteel and structured approach in his response. He countered Wright's remarks on the antecedence of Rasch's work, noting that Lord's early work was published in 1952, well before Rasch's work was published in 1960. He went on to stress that the model must fit the data, saying that when "Rasch's work became known in 1960, . . . Lord had already rejected that model due to its inferior fit to his data." He discussed the virtues of modeling item discrimination and guessing, noting the 80-year history of modeling discrimination and the improvements to model fit that these parameters afforded. Hambleton devoted significant time to reflecting on the practicality of implementation, noting that the software is sufficient and improving and that many testing programs are having success with multiparameter IRT. Finally, he closed by calling for pragmatism and making it clear that there is a place for many different models, including Rasch's, saying, "Rasch has an important role only when it

fits data well, and sample sizes are modest with no need for highly precise estimates” (Hambleton, 1992; Huang, 2015; Wright, 1992).

Neither Wright nor Hambleton was declared the winner, but with the benefits of hindsight we can see that the winner of the larger debate within the field would likely not have satisfied either side. There can be little doubt that among organizations using IRT for operational purposes the one-parameter models are the most popular. So, in a sense, Wright’s model was victorious. Importantly, however, Wright’s philosophy was not. Operational testing organizations are not choosing a Rasch model because it reflects their philosophical view of objective measurement; instead, they are acceding to Hambleton’s calls for pragmatism. The one-parameter model is popular because it accommodates modest sample sizes and is easy to implement for practitioners. In that way, it is safe to say that Hambleton’s philosophy won out, even if his preferred model did not.

Although Wright and Hambleton famously disagreed about the appropriate philosophical lens through which to view these models, they shared quite a bit in common. Both men made substantial contributions to the popularization of IRT. Both men wrote articles and influential books on their vision of IRT, and both were influential professors—Wright at the University of Chicago and Hambleton at the University of Massachusetts Amherst—who trained a generation of psychometricians in the use of IRT.

IRT Extensions

Almost as soon as the foundations of IRT were completed, researchers began developing extensions to both Rasch and multiparameter IRT to accommodate a wide variety of response data. This process was so productive with extensions building on extensions that van der Linden and Hambleton’s (1997) *Handbook of Modern Item Response Theory* presents no fewer than 27 models, including everything from polytomous extensions of the logistic models, to multidimensional extensions, to models incorporating response time.

One of the earliest IRT extensions is the result of work by Fumiko Samejima. Inspired by Lord’s early work in 1952, Samejima became interested in IRT and began considering the possibility of polytomous extensions. After graduating in 1965, Samejima took several positions in North America, first as a visiting research psychologist at ETS, where she met Lord. After a number of other short-term university positions, in 1973 she accepted a position at the University of Tennessee, Knoxville. Samejima published work describing both normal ogive and logistic versions of what she called the *graded response model* (Samejima, 1969, 1972). Her approach was to model the probability that a test taker of a given proficiency would score at or above a given level on a polytomously scored task (Barnhart, 2013).

The next major extension to IRT models was the *nominal response model* developed by Darrell Bock (1927–2021) in 1972. After graduating from the University of Chicago, Bock worked at the University of North Carolina at Chapel Hill, where he was briefly a colleague of Fumiko Samejima, before returning to teach at the University of Chicago in the same department as Ben Wright. Bock’s nominal response model was

conceptually similar to Samejima's heterogeneous graded response model but allowed for unordered categories. As a result, it was useful for the sort of unordered data that result from some psychological assessments as well as analyzing distractor patterns within educational assessments (Wainer & Robinson, 2006). David Thissen and Lynne Steinberg (1984) extended this latter use with their multiple-choice model. The multiple-choice model extends the nominal response model to support a guessing parameter to further support distractor analysis in education assessments.

The first polytomous extension from the Rasch perspective was provided by David Andrich (1978). After completing his undergraduate education in Australia, Andrich came to the United States to attend graduate school at the University of Chicago. Although Andrich worked with both Ben Wright and Darrell Bock, there is no doubt that he approached his work from a Rasch perspective. In fact, in 1977, Andrich spent 6 months working directly with Georg Rasch at the Danish Institute for Educational Research. Andrich's *rating scale model* estimates the probability that a person with a given ability would achieve each of the possible score points (Webster, 1998). The model, in addition to assuming that the slopes of the items are fixed, assumes that the threshold structure is consistent across all items on the exam. That is, the offsets between each score point and the nominal item difficulty are consistent across all items. Although this assumption may be difficult to satisfy for educational assessment, Andrich was primarily interested in Likert-type data, where the assumption may be more likely to be satisfied. In 1982, Geoff Masters, a fellow Australian and graduate of the University of Chicago, extended the rating scale model to make it more appropriate for educational assessments with the partial credit model. The partial credit model allowed the threshold structure of polytomous items to vary across the exam. This makes it possible for the difference between two specific score points to vary by item. As the name suggests, this can be applied in an educational context to provide some credit for a response that demonstrates partial mastery.

In 1992, the partial credit model was extended again by Eiji Muraki. Muraki, like Andrich and Masters, attended the University of Chicago for graduate school. Unlike Andrich and Masters, Muraki was advised by Darrell Bock and brought a multiparameter perspective to his extension of the partial credit model. In 1992, while studying at Chicago, Muraki published the *generalized partial credit model* that allowed the slope of the items to vary across the exam (Muraki, 1992). At roughly the same time, Wendy Yen, while working at CTB, independently developed the same model (Yen & Fitzpatrick, 2006). Although it was parameterized differently, Yen's two-parameter partial credit model and the generalized partial credit model are mathematically equivalent. This work produced independently by Muraki and Yen transformed the Rasch derived partial credit model to make it compatible with assessments that use a multiparameter IRT model.

Computerized Adaptive Testing

One prominent application of IRT is in the area of multistage and computerized adaptive testing, or what Lord originally referred to as “tailored testing.” Broadly speaking, these approaches route test takers to different items based on proficiency estimates that are produced iteratively as the test is administered. These approaches had been explored in a classical test theory context in the 1950s and 1960s with Angoff and Huddleston’s (1958) work on two-stage testing and Cleary et al.’s (1968) work on a variety of “programmed” testing methods, but these ideas needed an IRT framework to be practical. By the late 1960s, Lord had built on his earlier work on IRT to lay the theoretical foundation for adaptive testing and developed IRT-based methods of delivering both multistage tests with a router test and item-level computerized adaptive tests (Carlson & von Davier, 2017; Lord, 1968). By the early 1970s, Lord had gone further to address many of the practical challenges of adaptive testing, before the technology to deliver these tests was broadly available (Lord, 1970, 1971a, 1971b, 1971c, 1971d; Wainer, 2000).

Since the earliest conceptualizations of computerized adaptive testing there has been ongoing research on the best method for selecting items to be administered to each test taker in a manner that maximizes score precision while respecting nonstatistical constraints on item selection. One of the most popular solutions to this problem known as the *weighted deviation method* was introduced by Lord’s longtime collaborator, Martha Stocking (1942–2006). Lord hired Stocking in 1967 as a research scientist at ETS after she completed her master’s degree in statistics at Rutgers University (Bennett & von Davier, 2017; “Martha L. Stocking Swanson,” 2006). Although her most famous and broadly applied contribution is undoubtedly the *Stocking–Lord equating method* (Stocking & Lord, 1983), Stocking was also prolific in the area of computerized adaptive testing, where she published several important articles and reports on practical considerations in controlling items exposure within an adaptive framework (Stocking et al., 1991a, 1991b; Stocking & Lewis, 1998, 2000; Stocking & Swanson, 1993, 1998).

An alternative approach to selecting items within a computerized adaptive test, while respecting content constraints, was proposed by Howard Wainer in 1987 (Wainer & Kiely, 1987). Wainer was hired in 1980 as a principal research scientist at ETS, where he remained for 21 years (Robinson, 2005). While there, as a colleague of both Lord and Stocking, he developed the testlet-based approach to selecting items within an adaptive test. In the testlet-based approach, rather than selecting individual items, small groups of items, or *testlets*, are administered as a set. These predefined groups of items can be selected and reviewed by content experts to ensure they meet all blueprint constraints before the exam is administered. Although this provides greater control over test content, it provides less adaptability than approaches that select one item at a time (van der Linden, 2000).

A third approach to item selection within computerized adaptive tests was developed by Wim van der Liden (2000). This approach, known as a *shadow test*, attempted to

allow for selecting items one at a time, while ensuring that the complete form adheres to all the underlying content constraints. In a shadow test, after each item is administered, the selection algorithm selects an entire form for that test taker that includes all previously administered items and meets all blueprint constraints. The selection algorithm then selects the item from this shadow test that provides the most information given the test taker's current ability estimate. Given an appropriate item pool, this approach ensures that the test taker will receive an exam that meets the blueprint constraints and is tailored to their ability.

The foundational theory, model extensions, and practical developments including improved estimation procedures (see Bock, 1997; Luecht & Hambleton, 2021; and Thissen & Steinberg, 2020) provided a basis that has allowed IRT to become largely ubiquitous in educational measurement. The third and fourth editions of *Educational Measurement* included major chapters on the topic. The applications are now so diverse and widespread that the current edition recognizes that the topic can no longer be covered in a single chapter.

In the prior three sections we focused on theoretical and technical developments that have shaped the current practice of educational measurement. In the next section we return to a focus on the practice of assessment.

LARGE-SCALE TESTING

In the second section, we described how intelligence testing grew from its foundations in the work of Galton, Spearman, and the eugenics movement. Intelligence testing represents a critical part of the history of the practice of testing, but it is only one part of that history. There is an even longer history of standardized tests being used for selection, graduation, and the evaluation of individuals and schools. Over time, these efforts developed into large-scale testing programs. Much of this testing is targeted at achievement rather than intelligence.

The Chinese began using written examinations for civil service selection more than 3,000 years ago when candidates were tested in the “six arts”: music, archery, horsemanship, writing, arithmetic, and ceremonies of public life. Two thousand years ago, written tests were introduced in the “five studies”: civil law, military affairs, agriculture, revenue, and geography of the empire. Later tests emphasized remembering and interpreting the writings of Confucius. Even for these early examinations there was an understanding that standardization was critical (DuBois, 1970).

The practice—and history—of Chinese testing was known in England and influenced testing practice there. In 1833, standardized testing was instituted for the British Indian civil service. Over time, similar practices were adopted in the United States. In the 1860s, a bill was presented in Congress to institute a similar civil service system, but that bill was never enacted. Civil service testing was eventually adopted in the United States, but that adoption occurred in a piecemeal fashion. The various efforts impacted specific departments in the government and came and went over time as funding was

approved or ran out. A permanent civil service commission was finally established in 1883.

Although university-based tests can hardly be viewed as large-scale assessments, to provide context we note that examinations have played an important role in that context as well. Although oral examinations may have been part of Western education since the founding of the first universities, written tests also have a long tradition. The written examination that Karl Pearson completed to achieve the rank of third wrangler at Cambridge, the Mathematical Tripos, dates to the 18th century. By the late 19th century, written examinations were common in universities.

Horace Mann and Testing in Public Schools

Given the price of paper, it is not surprising that the adoption of written testing in grammar schools lagged behind that in universities. In the United States, grammar school students were typically assessed based on oral recitation in the first half of the 19th century. The drive to create standardized written student assessment in the United States was motivated by the desire to move beyond judgments about students to evaluation of the performance of schools. By early-21st-century standards, the suggestion that standardized testing within a single city should qualify as *large-scale testing* may seem quaint, but nonetheless large-scale testing in the United States appears to have its foundations in the tests administered in the Boston public schools at the instigation of Horace Mann (1798–1859).

Mann is an icon of American education; he fought for universal nonsectarian education, and in recognition of that effort, schools across the country are named in his honor. He had essentially no experience as a teacher and yet he had exceptionally strong opinions about teaching. Most important in the context of this chapter, Mann saw that written tests were not only useful to evaluate individuals; the tests could also be used to evaluate schools—and, by extension, school administrators.

Prior to the launch of these tests in 1845, Boston's schools were evaluated by local visiting committees. The individual schools were run by masters with little additional oversight. The visiting committees routinely found the schools to be performing well based on idiosyncratic and subjective criteria. Understandably, the masters wished to be left alone to run their schools as they thought best (Reese, 2013). By contrast, Mann and his allies believed standardized testing would bring a measure of quality control to a bureaucratically managed and centralized education system. They also believed that through this mechanism they could establish performance standards and systematize education, teacher evaluation, and pedagogy.

In late spring 1845, after months of controversy between Mann and the Boston masters, the first large-scale written examinations were administered to students in Boston. As Mann and his allies planned, the examination results were used both to understand student learning and to challenge the authority of the masters. School accountability testing arguably grew from Boston's 1845 examinations, and Mann's reform work in Boston started a trend toward standardized testing throughout the United States,

including most notably the New York Regents Examinations. The early Regents examinations were introduced in 1865 following the model of Boston's written examinations. The initial examinations were used for high school admissions and to provide an annual measure of student achievement (Phelps, 2007).

The Regents examinations were the vanguard of competitive testing in U.S. schools. As school enrollment increased through the late 1800s, testing expanded beyond classroom assessments to include standardized assessments for grade promotion, high school admission, and high school graduation. Leading educators of the time supported increasingly consolidated schools and standardized curricula as well as standardized examinations, which were thought to spur excellence in teachers and students while preparing students for challenges outside school settings (Reese, 2013).

The written examinations introduced by Mann and the New York Regents established the use of standardized written examinations for both selection and accountability in the United States. Intelligence testing then went on to influence standardized selection and accountability tests after World War I. This progression and interaction will be apparent when we discuss the evolution of the SAT. However, there is an independent strand in the history of testing that warrants attention; that strand represents the continued use of tests to assess achievement, both for accountability and to evaluate the progress of individual students. J. M. Rice is a noteworthy part of that strand.

Joseph Mayer Rice

Joseph Mayer Rice (1857–1934) remains relatively unknown, although his work laid the groundwork for research by better known leaders in educational research, including E. L. Thorndike and Terman (Graham, 1966). He trained and practiced as a pediatrician and, at some point, became interested in pedagogy. This interest motivated his travels in Europe to learn about the developing field of psychology and contemporary theories of pedagogy. After returning to the United States, he developed and implemented tests of spelling, mathematics, and language to evaluate the impact of the pedagogical methods used across schools in cities around the country. He published the results from these studies in the popular magazine *The Forum* (Rice, 1897a, 1897b); many of his original articles were included in his best-known work, *Scientific Management in Education* (Rice, 1913).²⁷

Rice succeeded in collecting large data sets to support his critique of the less-than-progressive educational practices that were prevalent at the time. His first study focused on spelling. In describing this work, Rice reported, "During this time three different tests were made; the number of children examined reaching nearly 33,000" (1897a, p. 164). The primary motivation of this data collection was to explore the relation between the amount of time used to drill spelling in the classroom and the outcomes measured by Rice's tests. Rice's contemporary, E. L. Thorndike, criticized Rice for being too quick to jump to speculative conclusions in interpreting his results (E. L. Thorndike, 1903). Nonetheless, it is reasonable to consider this data collection to be the beginning of the science of educational research. Stanley (1966) suggested that this was likely "the first

full-scale ‘experiment’ ever done in schools and published” (p. 135). It is similarly not difficult to find evidence to support the conclusion that Rice was a self-promoter who brought muckraking to the field of education (Graham, 1966).

School-Based Achievement Testing

In a previous section we discussed the rise of intelligence testing following World War I. At that same time, the use of achievement tests also grew dramatically. Students were often tested with both intelligence and achievement tests in an effort to evaluate whether their performance was in line with their potential.²⁸ The other SAT—the Stanford Achievement Test—was developed by Kelley, Ruch, and Terman in 1922. An advertisement for the test in Kelley (1927, end of volume) somewhat ironically describes the test as follows:

This new battery of achievement tests is designed to measure very thoroughly the knowledge and ability of pupils in school subjects in grades two through eight. It covers all the ground necessary to cover for ordinary purposes of educational testing.

The score in any subject is immediately comparable with the score in any other subject, and valid composite scores for any number of subjects taken together are readily obtainable. Age norms as well as grade norms make possible the derivation of a satisfactory educational quotient.

The battery included tests of arithmetic, reading, spelling, science information, and history and literature. A century after its introduction, the test—now in its 10th revision—is still in use.

In 1927, the Stanford Achievement Test was far from the only large-scale test available. Kelley (1927) listed 4 elementary school achievement batteries, as well as 12 elementary intelligence tests, 15 elementary reading tests, and 10 elementary spelling tests. Clearly, the decade after the First World War was a period of substantial growth for standardized testing.

Over time, other achievement tests were introduced. Perhaps the best known was the Iowa Test of Basic Skills, which continues to the present day as part of the Iowa Assessments. These tests supported a substantial testing industry, including for-profit corporations such as CTB/McGraw-Hill, Riverside, and the World Book Company. These large-scale achievement tests also provided motivation for improving testing procedures. Perhaps most notably, E. F. Lindquist—who was a major force behind the Iowa testing program—introduced the first optical scanner to score multiple-choice items.

So far, this section has focused on achievement tests. These tests were intended as accountability measures for schools, to assess individual progress, and as a means of evaluating the efficiency of differing pedagogical approaches. We now consider two tests used for college admissions: One had its origins in intelligence testing and the other in achievement testing.

The SAT

The group that developed and administered the Army Alpha test included luminaries of the science of intelligence testing such as Yerkes, Terman, Goddard, and Otis. The group also included Carl Brigham (1890–1943). The Army Alpha test was used to measure the “verbal ability, numerical ability, ability to follow directions, and knowledge of information” of military recruits. As we have seen, Army Alpha helped to launch the now-ubiquitous multiple-choice question; it also introduced intelligence testing on a large scale. Perhaps less well known is the fact that it was a steppingstone to the creation of the Scholastic Aptitude Test, the precursor to the current SAT. The connection between the two tests is straightforward; after the war, Brigham adapted the Army Alpha test to create the original SAT.

Established in 1900, the College Entrance Examination Board represented member colleges in the interest of developing uniform assessments (essays at the time) aligned with the curricula of schools that supplied graduates to elite institutions of higher education (Lemann, 2004). Previously, individual schools were responsible for developing their own admissions essays, and the College Board examinations introduced uniformity into the administration process, the tested content, and the scoring. The College Board essay tests were first administered in 1901 (Fuess, 1950). Though the goal of these examinations was standardization, College Board member schools expressed concerns in the 1920s about the subjectivity in scoring and the narrowing effect the exams had on high school curricula (Bennett, 2017).

These concerns led to the development of alternative assessment approaches, including multiple-choice achievement and aptitude tests, leading to the development and adoption of Brigham’s Scholastic Aptitude Test²⁹ in 1926 (Lemann, 2004).

Led by James Bryant Conant, then-president of Harvard University, in collaboration with Brigham and Henry Chauncey (then–assistant dean of admissions, Harvard University), the test was first employed to identify scholarship students based on their intellect, not their educational background. Harvard began using the test in 1934 to identify scholarship students, eventually requiring it of all applicants beginning in 1941. It became a requirement of College Board member schools in 1942.

The first version of the SAT, in use from 1926 to 1930, comprised nine subtests containing 315 multiple-choice questions; it was administered under highly speeded conditions, allowing 97 minutes for the full administration. The more familiar two-section format, testing verbal and mathematical aptitude, was introduced in 1930 (Lawrence et al., 2004).

As discussed earlier, intelligence testing rose to prominence during this same period. By 1932, a vast majority of large school systems in the United States were using intelligence testing to guide ability grouping (Haney, 1984). Despite the similarities in general appearance between intelligence tests and contemporaneous versions of the SAT, there is a marked tension between intelligence testing and the SAT. Throughout much of its existence, the developers of the SAT have argued that instead of measuring intelligence, versions of the SAT have been designed to measure aptitude and reasoning.

Currently, the College Board claims that its questions focus on skills that matter for college success.

ACT (Formerly American College Testing)

Although the SAT was the first large-scale college admissions test, it is no longer the largest. In response to the increasing number of college applicants following World War II and the federal initiatives of the Great Society to increase access to higher education, there was a high demand for standardized tools to support college admissions decisions. The SAT was commonly used in the northeastern states, but elsewhere in the country, states developed their own admissions tools, which increased the complexity of college application, particularly for out-of-state students (ACT, 2009). The American College Testing Program, also known as the ACT Test and ACT,³⁰ was an attempt to meet this need. The program had its roots in the Iowa Academic Meet, the first Iowa testing program for high school students.

The Iowa Academic Meet, or the “Brain Derby,” was developed in 1929 by E. F. Lindquist (editor of the first edition of *Educational Measurement*) under the direction of Professor Thomas Kirby (Holmgren, 2009). The Iowa Meet, in its early form, was used to identify outstanding scholars while raising the standards of instruction across Iowa high schools by testing every student (Lindquist, 1976, as referenced by Croft & Beard, 2021). This grew into the Iowa Assessments (previously known as the Iowa Testing Programs). Iowa Assessments were administered to students in Iowa beginning in 1935 and around the country beginning in the 1940s.

The ACT, cofounded by Lindquist and Ted McCarrell of the University of Iowa, was introduced in 1959 and was designed to assess a student’s general educational development. It incorporated multiple-choice questions that had previously appeared on the Iowa Test of Educational Development, allowing for rapid development (Croft & Beard, 2021). The original ACT had four sections—English, mathematics, social studies, and science—and was delivered in 3 hours (Croft & Beard, 2021; see ACT, 2009, for further details).

The ACT was originally tied to the assessment of knowledge, in contrast to the SAT’s focus on aptitude. Although this historical distinction exists, it would appear to be of little practical importance given the high correlation that exists between the tests—typically reported to be approximately .90—and the existence of widely used concordance tables that allow the scores to be used interchangeably (College Board & ACT, 2018; Perez, 2002).

In the previous pages, we considered tests designed for accountability, tests designed as a metric of individual student achievement, and tests for selection. So far, the tests we have discussed have one thing in common: They produce an interpretable score for each test taker. We conclude this section by considering a test explicitly designed so that individual scores cannot be produced.

NAEP

In 1867, Congress established a Department of Education. Among the purposes of the department was “collecting such statistics and facts as shall show the condition and progress of education in the several States and Territories” (An Act to Establish a Department of Education, 1867, Sec. 1). In essence, the department was to produce a national assessment of educational progress. This did not happen; the Department of Education was reduced to an office and then made part of the Department of the Interior.

Nearly 100 years later, the idea of a national assessment of educational progress was revived. (See also Ho & Polikoff, this volume, for a review of the history and current use of NAEP for accountability purposes.) In 1964, the test was funded by the Carnegie Corporation. The Education Commission of the States subsequently took over administration with funding from the Carnegie Corporation and then with increasing funding from the U.S. Office of Education. Congress took over all funding for NAEP in 1972 and moved NAEP to the National Institute of Education in 1978.

From the outset there was an effort to make NAEP an outstanding example of measurement science. The office of education—and later the Department of Education—contracted and consulted with the most respected organizations in the country, including ETS, the American Council on Education, American College Testing (ACT), Human Resources Research Organization, and the American Institutes for Research. This led to innovation and the application of state-of-the-art methodology.

Ralph Tyler (1902–1994), a contributor to the first edition of *Educational Measurement*, was an early advisor on the project. It was Tyler’s recommendation to base the national assessment on a systematic sampling strategy rather than a more exhaustive program of testing. Sampling reduced cost, but also ruled out the possibility of producing scores for individuals, which was prohibited by the legislation.

In 1983, the sophistication of the NAEP administration again took a major step forward. ETS was awarded a contract by the Department of Education to design and conduct the examination. ETS proposed numerous changes (Messick et al., 1983). They included a more sophisticated sampling design as well as the introduction of IRT to support test construction and to carry out linking across test forms and across years.

As we noted previously, NAEP has been an example of excellence and innovation in assessment science since the 1960s. In the IRT chapter from the fourth edition of *Educational Measurement*, Yen and Fitzpatrick (2006) described NAEP as “the most famous example of a matrix-sampled test.” They go on to say that “NAEP is also notable for the use of IRT in combination with advances in missing data technology and hierarchical analyses to estimate population characteristics without estimating individual examinee scores” (p. 145). NAEP was referenced in five separate chapters in the 1989 edition of *Educational Measurement* and in eight separate chapters of the fourth edition (Brennan, 2006).³¹

As this section suggests, much of the development of large-scale testing has happened in the United States, but in recent decades these tests have become important around the globe. For example, in China, the National College Entrance Examination, which determines eligibility for postsecondary education, is taken by approximately 10 million students each year. Because of the impact this test can have on future career opportunities, students spend years preparing for it (Larmer, 2014). High school graduation and university selection tests are also an established part of education in many other countries (see Clauser & Margolis, 2023). Beginning in the 1990s, three large-scale achievement tests have also been administered internationally in multiple languages: these include the Trends in International Mathematics and Science Study (TIMSS, first administered in 1996), the Programme for International Student Assessment (PISA, first administered in 2000), and the Progress in International Reading Literacy Study (PIRLS, first administered in 2001). These international tests are discussed in detail in Chapter 20 of this volume by Braun and Kirsch.

THE HISTORY OF EDUCATIONAL MEASUREMENT AS REFLECTED IN FIVE EDITIONS OF *EDUCATIONAL MEASUREMENT*

The multiple editions of *Educational Measurement* now span more than 7 decades (Brennan, 2006; Lindquist, 1951; Linn, 1989; R. L. Thorndike, 1971), and in this final section we consider how these volumes fit into the history of *Educational Measurement*. Since the first edition was published, it has been a highly regarded standard reference. The volumes can fairly be seen as representing the state of the art for educational measurement, but in general they do not represent the cutting edge (for example, the 1971 edition included only passing mention of the relatively new technology represented by IRT, although it had been thoroughly described by Birnbaum in Lord and Novick's 1968 text and by Rasch in 1960).

Although there are lead authors responsible for the chapters in each of the volumes, the works have been intentionally collaborative. The Lindquist volume lists collaborators along with the lead authors on the title page of each chapter. Subsequent volumes all highlight the reviewers for each chapter. Although there is sometimes disagreement within the field about some of the views expressed, the resulting chapters are at least intended to make a consensus statement about their specific topics. As such, changes across chapters from one edition to the next provide a useful view of how the field has evolved. In reviewing the third edition, Cronbach (1989b) commented on the increased length of each subsequent volume, the topics covered, and the level of mathematical sophistication required of the reader. In addition to these changes across editions, we also consider how the authors have changed over 70 years.

A synoptic view of the four earlier volumes provides considerable insight. Each of the volumes has a section on measurement theory. In the third and fourth editions, this is broadened to include "theory and principles." In the first two editions, this section is

placed after the more practical sections on test development and administration. In the subsequent editions, theory is (literally) given priority by being placed at the beginning of the volume. This seems to reflect a change in the intended audience of the volumes; it may also reflect a change in the field, with higher expectations that practitioners have advanced training. In a sense, these subsequent volumes demonstrate that the place of the measurement profession was now established.

One interesting way to look across the editions is to consider how the content has changed in subsequent chapters on the same topic. Examining the parallel chapters provides a detailed outline of historical changes in thinking within the field. For example, the chapters on validity allow the reader to see how the concept grew from validity as correlation and prediction (in Cureton's 1951 chapter) to include construct validity (as described by Cronbach, 1971). Messick (1989) then broadened the view of construct validity and gave particular attention to consequences and Kane (2006) formalized the idea of validity as a structured argument in support of the inferences we make based on test scores. Green (2008) pointed out that in the fourth edition correlation and prediction are almost absent from the presentation of validity.

Kolen (2021), in discussing the history of equating, made a similar case for the historical value of the early editions of *Educational Measurement*. He described Flanagan's (1951) chapter as the first comprehensive description of equating and discussed how Angoff (1971)—writing in the second edition—reconsidered some of the basic assumptions of equating (e.g., the idea that equating should be sample invariant). As is the case with validity theory, changes in practice continue to be documented in the subsequent editions as equal percentile and IRT-based approaches receive more attention.

In addition to considering how similar content has evolved across parallel chapters in the editions of *Educational Measurement*, it is interesting to note how the content has changed across volumes. In all five editions the section on theory includes chapters on validity, reliability, and equating. The first two editions also included chapters related to theory, which were dropped from subsequent editions. Both the first and the second editions included chapters on approaches for considering multiple measurements (i.e., batteries and profiles). As less attention has been given to intelligence testing, these sections appear to have been considered less relevant. These two early editions also included chapters on the *nature of measurement*; this topic was not included in the third or fourth editions but has a reprise in the current edition.

Not surprisingly, the section on theory and principles in subsequent editions also includes chapters that were not included in the first two volumes. Three chapters in the Linn (1989) edition clearly reflect trends in the field that have received considerable attention since the second edition was published: IRT, cognitive psychology, and bias in testing. As Cronbach (1989a) pointed out, although by 1971 there had already been a long presentation of IRT by Birnbaum and a book on computer-assisted testing, there is almost no mention of IRT in the second edition. The attention given to this area since the 1970s is reflected by the chapters included in the third and fourth editions.

Similarly, the controversy over racial differences in IQ scores, concern for equity and fairness, and the increased attention given to consequences as an integral part of test validity led to a very significant focus on test bias. This is reflected in the chapter on bias in the third edition. Finally, chapters on cognitive psychology in the third and fourth editions reflect the extent to which measurement experts searched for a theory-based foundation for assessment practice.

In addition to considering the content of the editions, it is interesting to consider how the contributors have changed over time. In the first three editions, the majority of the authors were affiliated with universities (57%, 61%, and 69%, respectively). In the two most recent editions, fewer than 50% of the authors had university affiliations; most were associated with nonprofit organizations. Relatively few of the contributors to any of the editions have come from for-profit testing companies.

The authors who have made contributions to *Educational Measurement* have also changed in another important way over time. The list of authors in the first two editions included only one woman.³² The percentage of women contributing to the volumes has substantially increased with the subsequent editions. Across the four earlier editions, the percentage of women authors was 4, 0, 27, and 35, respectively. In the current edition more than 40% of the authors are women.

SOME FINAL THOUGHTS

In this chapter we have described some of the critical themes and events in the history of educational measurement. We believe that understanding the history of our field is important. To use a phrase that Howard Wainer (2005, p. 118) attributed to Aristotle, “We understand best those things we see grow from their very beginnings.” If we wish to understand the field of educational measurement, we would be well served by understanding how it developed. That history tells us something of what has been accomplished since the time of Francis Galton. It is easy to view the development of correlation by Galton and Pearson; the adoption of those methods for correlational psychology by Spearman; the extension of Spearman’s formulas by Kelley, Kuder and Richardson, and Cronbach; and finally the development of generalizability theory as an inevitable evolution. The history as presented in this chapter documents that evolution, but it also makes clear that there was nothing inevitable about it. The dominant approach to understanding reliability could instead have been based on Edgeworth’s generalizability-like framework—which was published before Galton’s introduction of the correlation coefficient. The evolution of test theory, like Darwinian evolution, has not been a progression toward some abstract perfection; it simply reflects the relative usefulness of different approaches given the environment in which they exist. Spearman’s work supported an exploration of the nature of intelligence, and that exploration was well suited to an environment shaped by the widely popular eugenics movement.

These beginnings have had a lasting impact on both the theory and the practice of testing. As we discussed in the second section, the Army Alpha test brought to prominence large-scale intelligence testing using multiple-choice questions. The focus on intelligence testing had a major impact on testing practice—up to and including early versions of the SAT, which were modeled after the Army Alpha test; the general format represented by these tests (if not the content) has remained dominant for over a century.

The early prevalence of intelligence testing also created a close link between testing theory and practice and the eugenics movement. This link helps to explain why the public may not trust experts who suggest that they can benefit society by measuring minds. The eugenics movement caused horrific suffering both in Europe and in the United States. Many of the great thinkers in statistics and educational measurement were closely aligned with that movement. Names like Galton, Pearson, Spearman, Fisher, Goddard, and Kelley represent a very partial list of eugenicists associated with educational measurement.

In understanding the context that has produced the practice of testing in the 21st century, it is also important to remember the political forces that have shaped testing. At least as far back as the 1860s, the U.S. Congress has made political decisions about whether to support—or to withhold support from—testing. The civil service commission and the army testing program during World War I are early examples, but the testing mandates put in place to evaluate the American education system and to provide accountability (such as NAEP) have likely done even more to shape the field of educational measurement.

Another example of the importance of context arises in the sections on scaling and the development of IRT. These sections are closely linked and both remind us that beliefs about the nature of educational measurement have not only changed over time, but also varied across researchers at the same time. Fechner constructed a scale that reflected his understanding of measurement from the study of physics. Binet, by contrast, rejected the link between psychological and physical measurement, stating, “The scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not super-posable, and therefore cannot be measured as linear surfaces” (Binet & Simon, 1916, p. 40). This split between pragmatism and philosophical beliefs about the nature of measurement resulted in high-profile disagreements between the adherents of multiparameter IRT and Rasch measurement. That debate continues in more recent works such as those by Mari et al. (2021) and Irribarra (2021). Understanding how these differences of opinion have persisted should help the reader to hold these disagreements in perspective.

Each of the sections in this chapter provides an example of why we believe that the theory and practice of educational measurement needs to be understood in their historical context. It is easy to read a text on classical test theory or IRT and come away with the impression that they are inevitable truths describing immutable relationships. These theories are simply tools that have proven useful in specific contexts. These tools are valuable (and the present authors have spent decades trying to mas-

ter their use), but as the questions we ask change, the tools may need to change as well. Historical context provides an understanding of how and when those changes become critical.

In the end, it is not our intention to tell the reader how to interpret the history of our field. We certainly do not have answers to the question of what lessons should be learned from that history. But we do believe that leaders in educational measurement will understand the field better when they understand how it has grown from its beginnings.

ACKNOWLEDGMENTS

One of the great challenges of writing this chapter is the great expanse represented by the history of educational measurement. It would be impossible for any set of authors to know it all. We are, therefore, humbled by the tremendous support we have received throughout the process. First, we would like to thank Linda Cook and Mary Pitoniak for entrusting us with this work, as well as for their leadership and thoughtful suggestions throughout. We would also like to thank Howard Wainer, Mark Reckase, and Kadriye Ercikan, who acted as reviewers and collaborators on the chapter, for their multiple reviews filled with many valuable suggestions. Sometimes these suggestions corrected errors, but more importantly, they led us to include historical figures and events that we may otherwise have overlooked. Their contributions have improved both the content and the form of the chapter.

REFERENCES

- Abelson, A. R. (1911). The measurement of mental ability of "backward" children. *British Journal of Psychology*, 4(3-4), 268-314. <https://doi.org/10.1111/j.2044-8295.1911.tb00047.x>
- ACT. (2009). *ACT: The first fifty years, 1959-2009*.
- An Act to Establish a Department of Education, Pub. L. No. 39-73, 14 Stat. 434 (1867). <https://www.docsteach.org/documents/document/act-of-march-2-1867-public-law-3973-14-stat-434-which-established-the-department-of-education>
- Aikens, H. A., Thorndike, E. L., & Hubbell, E. (1902). Correlations among perceptive and associative processes. *Psychological Review*, 9(4), 374-382. <https://doi.org/10.1037/h0070072>
- Andersen, E. B., & Olsen, L. W. (2001). The life of Georg Rasch as a mathematician and as a statistician. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 3-24). Springer. https://doi.org/10.1007/978-1-4613-0169-1_1
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573. <https://doi.org/10.1007/BF02293814>
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). American Council on Education.

- Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-stage test system for the College Board Scholastic Aptitude Test* (Statistical Report SR-58-21). ETS.
- APF Gold Medal Awards and Distinguished Teaching of Psychology Award. (1993). *American Psychologist*, 48(7), 717–725. <https://doi.org/10.1037/h0090746>
- Bagley, W. C. (1901). On the correlation of mental and motor ability in school children. *The American Journal of Psychology*, 12(2), 193–205. <https://doi.org/10.2307/1412533>
- Barnard, G. A., & Godambe, V. P. (1982). Memorial article: Allan Birnbaum 1923–1976. *The Annals of Statistics*, 10(4), 1033–1039. <https://doi.org/10.1214/aos/1176345968>
- Barnhart, R. (2013). *Fumiko Samejima (b. 1930)*. Society for the Psychology of Women. <https://www.apadivisions.org/division-35/about/heritage/fumiko-samejima-biography>
- Bennett, R. E. (2017). What does it mean to be a nonprofit educational measurement organization in the twenty-first century? In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-319-58689-2_1
- Bennett, R. E., & von Davier, M. (Eds.). (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS*. Springer. https://doi.org/10.1007/978-3-319-58689-2_1
- Binet, A., & Fere, C. (1887). *Le magnetisme animal*. Alcan.
- Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet–Simon Scale)* (E. S. Kite, Trans.). Williams and Williams Co. <https://doi.org/10.1037/11069-000>
- Birnbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems* (Report No. 58-16). USAF School of Aviation Medicine.
- Birnbaum, A. (1958a). *Further considerations of efficiency in tests of a mental ability* (Report No. 17). USAF School of Aviation Medicine.
- Birnbaum, A. (1958b). *On the estimation of mental ability* (Report No. 15). USAF School of Aviation Medicine.
- Birnbaum, A. (1967). *Statistical theory for logistic mental test models with a prior distribution of ability* (ETS Research Bulletin No. RB-67-12). ETS. <https://doi.org/10.1002/j.2333-8504.1967.tb00363.x>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison–Wesley.
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 16(4), 21–33. <https://doi.org/10.1111/j.1745-3992.1997.tb00605.x>
- Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). Prentice Hall.
- Brennan, R. L. (1983). *Elements of generalizability theory*. ACT.

- Brennan, R. L. (2001a). *Generalizability theory*. Springer-Verlag. <https://doi.org/10.1007/978-1-4757-3456-0>
- Brennan, R. L. (2001b). *Manual for mGENOVA*. University of Iowa, Iowa Testing Programs.
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). American Council on Education; Praeger.
- Brennan, R. L. (2021). Generalizability theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 206–231). Routledge. <https://doi.org/10.4324/9780367815318-10>
- Briggs, D. C. (2021). A history of scaling and its relationship to measurement. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 263–291). Routledge. <https://doi.org/10.4324/9780367815318-12>
- Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge. <https://doi.org/10.1201/9780429275326>
- Brigham, C. C. (1923). *A study of American intelligence*. Princeton University Press.
- Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review*, 37(2), 158–165. <https://doi.org/10.1037/h0072570>
- Brooks, J. L. (1984). *Just before the origin: Alfred Russel Wallace's theory of evolution*. Columbia University Press.
- Brown, W. (1910a). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296–322. <https://doi.org/10.1111/j.2044-8295.1910.tb00207.x>
- Brown, W. (1910b). *The use of the theory of correlation in psychology*. Cambridge University Press.
- Brown, W. (1911). *The essentials of mental measurement*. Cambridge University Press.
- Bunch, M. B. (2021). The role of the federal government in shaping educational testing policy and practice. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 65–86). Routledge. <https://doi.org/10.4324/9780367815318>
- Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245–314). Macmillan.
- Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology*, 8(2), 103–118. <https://doi.org/10.1111/j.2044-8317.1955.tb00325.x>
- Capshew, J. H. (1999). *Psychologists on the march: Science, practice, and professional identity in America, 1929–1969*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511572944>
- Carlson, J. E., & von Davier, M. (2017). Item response theory. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 133–178). Springer. https://doi.org/10.1007/978-3-319-58689-2_1

- Cherry, K. (2023, September 28). *Edward Thorndike's contributions to psychology*. Very Well Mind. <https://www.verywellmind.com/edward-thorndike-biography-1874-1949-2795525>
- Clauser, B. E. (2021). A history of classical test theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 157–180). Routledge. <https://doi.org/10.4324/9780367815318-8>
- Clauser, B. E., & Margolis, M. J. (2023). Past, present, and future of educational measurement. In R. J. Tierney, F. Rizvi, & K. Erkican (Eds.), *International encyclopedia of education: Vol. 14. Quantitative research and educational measurement* (pp. 1–14). Elsevier. <https://doi.org/10.1016/b978-0-12-818630-5.10001-6>
- Clark, C. L. (Ed.). (1976). *Proceedings of the first conference on computerized adaptive testing*. U.S. Government Printing Office.
- Cleary, T. A., Linn, R. L., & Rock, D. A. (1968). An exploratory study of programmed tests. *Educational and Psychological Measurement*, 28(2), 345–360. <https://doi.org/10.1177/001316446802800212>
- Clifford, G. J. (n.d.). *Edward L. Thorndike (1874–1949): The man and his career, a psychology for educators, education as specific habit formation*. Education Encyclopedia. <https://education.stateuniversity.com/pages/2509/Thorndike-Edward-L-1874-1949.html>
- College Board & ACT. (2018). *Guide to the 2018 ACT/SAT concordance*. <https://satsuite.collegeboard.org/media/pdf/guide-2018-act-sat-concordance.pdf>
- Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 27(4), 907–949. <https://doi.org/10.1214/aoms/1177728067>
- Croft, M., & Beard, J. J. (2021). Development and evolution of the SAT and ACT. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 22–41). Routledge. <https://doi.org/10.4324/9780367815318-2>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.
- Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, 30(1), 1–14. <https://doi.org/10.1037/0003-066X.30.1.1>
- Cronbach, L. J. (1989a). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Lawrence Erlbaum.
- Cronbach, L. J. (1989b). Review of *Educational Measurement*, third edition. *Educational Measurement: Issues and Practice*, 8(4), 22–25. <https://doi.org/10.1111/j.1745-3992.1989.tb00339.x>
- Cronbach, L. J. (1991). Methodological studies—a personal perspective. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 385–400). Lawrence Erlbaum.

- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2), 137–163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>
- Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391–418. <https://doi.org/10.1177/0013164404266386>
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.
- Darwin, C. (1859). *On the origin of species*. John Murray.
- Darwin, C. (1869, December 23). [Letter to Francis Galton]. Darwin Correspondence Project (Letter no. 7032). <https://www.darwinproject.ac.uk/letter/DCP-LETT-7032.xml>
- Dorans, N. J. (2004). *A conversation with Ledyard R. Tucker*. ETS. <http://www.ets.org/Media/Research/pdf/TUCKER.pdf>
- Dr. Paul Lazarsfeld dies. (1976, September 1). *The New York Times*. <https://www.nytimes.com/1976/09/01/archives/dr-paul-lazarsfeld-dies-sociologist-at-columbia.html>
- DuBois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.
- Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51, 346–368. <https://doi.org/10.1177/014662168400800104>
- Education Amendments of 1974, Pub. L. 93-380, 88 Stat. 484 (1974). <https://www.govinfo.gov/content/pkg/COMPS-727/pdf/COMPS-727.pdf>
- Engelhard, G., Jr. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, 8(1), 21–38. <https://doi.org/10.1177/014662168400800104>
- Fechner, G. T. (1966). *Elements of psychophysics* (H. E. Adler, Trans.; D. H. Howes & E. G. Boring, Eds.; Vol. 1). Rinehart & Winston. (Original work published 1860)
- Fisher, I. (1923, December 21). Letter to Truman Kelley from the Eugenics Committee of the United States of America. Harvard University Archives: Truman Kelley Papers.
- Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.
- Flanagan, J. C. (1939). *The cooperative achievement tests. A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores*. American Council on Education Cooperative Test Service.
- Flanagan, J. C. (1942). A study of factors determining family size in a selected professional group. *Genetic Psychology Monographs*, 25, 3–99.

- Flanagan, J. C. (Ed.). (1948). *The Aviation Psychology Program in the Army Air Forces* (Vol. 1). U.S. Government Printing Office. <https://doi.org/10.1037/e614382011-001>
- Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). American Council on Education.
- Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51(4), 327–358. <https://doi.org/10.1037/h0061470>
- Fuess, C. M. (1950). *The College Board: Its first fifty years*. Columbia University Press.
- Galton, F. (1853). *Tropical South Africa*. John Murray.
- Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. Macmillan. <https://doi.org/10.1037/13474-000>
- Galton, F. (1875). Statistics by intercomparison with remarks on the law of frequency of error. *Philosophical Magazine*, 49(322), 33–46. <https://doi.org/10.1080/14786447508641172>
- Galton, F. (1884). *Anthropometric laboratory*. William Clowes and Sons.
- Galton, F. (1886a). Family likeness in eye-colour. *Proceedings of the Royal Society*, 40(242–245), 402–417. <https://doi.org/10.1098/rspl.1886.0058>
- Galton, F. (1886b). Family likeness in stature. *Proceedings of the Royal Society*, 40(242–245), 40–72. <https://doi.org/10.1098/rspl.1886.0009>
- Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society*, 45(273–279), 135–145. <https://doi.org/10.1098/rspl.1888.0082>
- Galton, F. (1889). *Natural inheritance*. Macmillan. <https://doi.org/10.5962/bhl.title.32181>
- Galton, F. (1892). *Finger prints*. Macmillan.
- Galton, F. (1893). *Decipherment of blurred finger prints*. Macmillan.
- Galton, F. (1895). *Finger print directories*. Macmillan.
- Galton, F. (1908). *Memories of my life*. Methuen. <https://doi.org/10.5962/bhl.title.28398>
- Galton, F. (1909). *Essays in eugenics*. Eugenics Education Society. <https://doi.org/10.5962/bhl.title.168760>
- Gillham, N. W. (2001). *The life of Sir Francis Galton: From African exploration to the birth of eugenics*. Oxford University Press. <https://doi.org/10.1093/oso/9780195143652.001.0001>
- Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30, 395–418. <https://doi.org/10.1007/BF02289531>
- Goddard, H. H. (1912). *The Kallikak family: A study in the heredity of feeble-mindedness*. Macmillan.
- Gottfredson, L. S. (1994, December 13). Mainstream science on intelligence. *Wall Street Journal*, A18.

- Gottfredson, L. S. (2018). G theory: How recurring variation in human intelligence and the complexity of everyday tasks create social structure and the democratic dilemma. In R. J. Sternberg (Ed.), *The nature of human intelligence* (pp. 130–151). Cambridge University Press.
- Gould, S. J. (1996). *The mismeasure of man*. Norton & Company.
- Graham, P. A. (1966). Joseph Mayer Rice as a founder of the progressive education movement. *Journal of Educational Measurement*, 3(2), 129–133. <https://doi.org/10.1111/j.1745-3984.1966.tb00868.x>
- Green, B. F. (1951a). *Latent class analysis: A general solution and an empirical evaluation* (ETS Research Bulletin No. RB-51-15). ETS. <https://doi.org/10.1002/j.2333-8504.1951.tb00215.x>
- Green, B. F. (1951b). A general solution for the latent class model of latent structure analysis. *Psychometrika*, 16, 151–166. <https://doi.org/10.1007/BF02289112>
- Green, B. F. (2008). Book review: *Educational measurement* (4th ed.). *Journal of Educational Measurement*, 45(2), 195–200. <https://doi.org/10.1111/j.1745-3984.2008.00060.x>
- Gulliksen, H. (1936). The content reliability of a test. *Psychometrika*, 1, 186–194.
- Gulliksen, H. (1950). *Theory of mental tests*. Wiley. <https://doi.org/10.1037/13240-000>
- Haley, D. C. (1952). *Estimation of dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Applied Mathematics and Statistics Laboratory, Stanford University.
- Hambleton, R. K. (1992). Hambleton's 9 theses. *Rasch Measurement Transactions*, 6(2), 215–217. <https://www.rasch.org/rmt/rmt62d.htm>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff Publishing. <https://doi.org/10.1007/978-94-017-1988-9>
- Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research*, 54(4), 597–654. <https://doi.org/10.3102/00346543054004597>
- Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology*, 5(1), 51–84. <https://doi.org/10.1111/j.2044-8295.1912.tb00055.x>
- Herrnstein, R. J. (1971). I.Q. *Atlantic Monthly*, 228(3), 43–64.
- Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. Simon & Schuster.
- Hilgard, E. R. (1989). The early years of intelligence measurement. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 7–28). University of Illinois Press.
- Holahan, C. K., Sears, R. R., & Cronbach, L. J. (1995). *The gifted group in later maturity*. Stanford University Press.
- Holmgren, D. (2009). Lindquist, Everet Franklin. In D. Hudson, M. Bergman, & L. Horton (Eds.), *The biographical dictionary of Iowa*. University of Iowa Press Digital Editions.

- Holzinger, K. J., & Clayton, B. (1925). Further experiments in the application of Spearman's prophecy formula. *Journal of Educational Psychology*, 16(5), 289–299. <https://doi.org/10.1037/h0075199>
- Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153–160. <https://doi.org/10.1007/BF02289270>
- Huang, W. (2015, December 15). *Benjamin Wright, renowned psychometrician, 1926–2015*. UChicago News. <https://news.uchicago.edu/story/benjamin-wright-renowned-psychometrician-1926-2015>
- Irribarra, D. T. (2021). *A pragmatic perspective of measurement*. Springer. <https://doi.org/10.1007/978-3-030-74025-2>
- Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 39(1), 1–123.
- Jensen, A. R. (1980, April). Letter to Janice Scheuneman. Collection of the authors.
- Jensen, A. R. (1998). *The G factor: The science of mental ability*. Praeger.
- Jones, L. V. (2007). Some lasting consequences of the U.S. psychology programs in World Wars I and II. *Multivariate Behavioral Research*, 42(3), 593–608. <https://doi.org/10.1080/00273170701382542>
- Jones, L. V., & Olkin, I. (2004). *The nation's report card: Evolution and perspectives*. Phi Delta Kappa Educational Foundation.
- Kane, M. T. (2002). Inferences about variance components and reliability—generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, 39(2), 165–181. <https://doi.org/10.1111/j.1745-3984.2002.tb01141.x>
- Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.
- Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research*, 3(5), 370–379. <https://doi.org/10.1080/00220671.1921.10879169>
- Kelley, T. L. (1923). *Statistical method*. Macmillan.
- Kelley, T. L. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. *Journal of Educational Psychology*, 16, 300–303. <https://doi.org/10.1037/h0073506>
- Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book.
- Kelly, F. J. (1915). *The Kansas Silent Reading Test*. Kansas State Printing Plant.
- Kendall, M. G. (1968). Studies in the history of probability and statistics. XIX Francis Ysidro Edgeworth, 1845–1926. *Biometrika*, 55(2), 269–275.
- Kolen, M. (2021). History of test equating methods and practices through 1985. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 318–342). Routledge. <https://doi.org/10.4324/9780367815318-14>
- Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of reliability. *Psychometrika*, 2, 151–160. <https://doi.org/10.1007/BF02288391>

- Larmer, B. (2014, December 31). Inside a Chinese test-prep factory. *New York Times Magazine*. <https://www.nytimes.com/2015/01/04/magazine/inside-a-chinese-test-prep-factory.html>
- Lawrence, I., Rigol, G., Van Essen, T., & Jackson, C. (2004). A historical perspective on the content of the SAT. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 57–74). Routledge Falmer. <https://doi.org/10.4324/9780203463932>
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Princeton University Press.
- Lemann, N. (2004). A history of admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 5–14). Routledge Falmer. <https://doi.org/10.4324/9780203463932>
- Levy, R., & Mislevy, R. J. (2021). A history of Bayesian inference in educational measurement. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 292–317). Routledge. <https://doi.org/10.4324/9780367815318-13>
- Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.
- Lindquist, E. F. (1976, October 5). *Interview by James Beilman with Dr. Everet F. Lindquist* (Interview transcript No. XXII). University of Iowa Oral History Project, 1976–1977.
- Linn, R. L. (Ed.). (1989). *Educational measurement* (3rd ed.). American Council on Education; Macmillan.
- Lippmann, W. (1922a). The mental age of Americans. *The New Republic*, 32(412), 213–215.
- Lippmann, W. (1922b). The mental age of Americans (Part II): Mystery of the “A” men. *The New Republic*, 32(413), 246–248.
- Lippmann, W. (1922c). The mental age of Americans (Part III): Reliability of the intelligence tests. *The New Republic*, 32(414), 275–277.
- Lippmann, W. (1922d). The mental age of Americans (Part IV): Abuse of the tests. *The New Republic*, 32(415), 297–298.
- Lippmann, W. (1922e). The mental age of Americans (Part V): Tests of hereditary intelligence. *The New Republic*, 32(416), 328–330.
- Lippmann, W. (1922f). The mental age of Americans (Part VI): A future for the tests. *The New Republic*, 33(417), 9–11.
- Lippmann, W. (1923). The great confusion. *The New Republic*, 33(422), 145–146.
- Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Psychometric Corporation.
- Lord, F. M. (1954). Scaling. *Review of Educational Research*, 24, 375–393. <https://doi.org/10.2307/1169042>

- Lord, F. M. (1968). *Some theory for tailored testing* (ETS Research Bulletin No. RB-68-38). ETS. <https://doi.org/10.1002/j.2333-8504.1968.tb00562.x>
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139–183). Harper & Row.
- Lord, F. M. (1971a). Robbins–Monro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1), 3–31. <https://doi.org/10.1177/001316447103100101>
- Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement*, 8(3), 147–151. <https://doi.org/10.1111/j.1745-3984.1971.tb00918.x>
- Lord, F. M. (1971c). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66(336), 707–711. <https://doi.org/10.1080/01621459.1971.10482333>
- Lord, F. M. (1971d). A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 31(4), 805–813. <https://doi.org/10.1177/001316447103100401>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison–Wesley.
- Lovie, P., & Lovie, A. D. (1996). Charles Edward Spearman, F.R.S. (1863–1945). *Notes and Records of the Royal Society of London*, 50(1), 75–88. <https://doi.org/10.1098/rsnr.1996.0007>
- Lovie, S., & Lovie, P. (2010). Commentary: Charles Spearman and correlation: A commentary on “the proof and measurement of association between two things.” *International Journal of Epidemiology*, 39(5), 1151–1153. <https://doi.org/10.1093/ije/dyq183>
- Luecht, R. M., & Hambleton, R. K. (2021). Item response theory: A historical perspective and brief introduction to applications. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 232–262). Routledge. <https://doi.org/10.4324/9780367815318-11>
- Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: Developing a shared concept system for measurement*. Springer. <https://doi.org/10.1007/978-3-030-65558-7>
- Martha L. Stocking Swanson. (2006, December 29). *The Times of Trenton*. <https://obits.nj.com/us/obituaries/trenton/name/martha-swanson-obituary?id=14042800>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174. <https://doi.org/10.1007/BF02296272>
- McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, 152, Article 107346. <https://doi.org/10.1016/j.measurement.2019.107346>
- Mendel, J. G. (1866). Versuche über pflanzen-hybriden [Experiments on plant hybrids]. *Verhandlungen des Naturforschenden Vereines in Brünn, Bd. IV*, 3–47. <https://doi.org/10.5962/bhl.title.61004>

- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955–966. <https://doi.org/10.1037/0003-066X.30.10.955>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education; Macmillan.
- Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress reconsidered: A design for a new era* (Report No. NAEP-83-01). ETS.
- Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511490040>
- Minton, H. L. (1990). Lewis M. Terman and mental testing: In search of the democratic ideal. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 95–112). Rutgers University Press.
- Morant, G. M., & Welch, B. L. (1939). *A bibliography of the statistical and other writings of Karl Pearson*. Cambridge University Press.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Department of Education.
- Otis, A. S. (1918a). An absolute point scale for the group measurement of intelligence. Part 1. *Journal of Educational Psychology*, 9(5), 239–261. <https://doi.org/10.1037/h0072885>
- Otis, A. S. (1918b). An absolute point scale for the group measurement of intelligence. Part 2. *Journal of Educational Psychology*, 9(6), 333–348. <https://doi.org/10.1037/h0074753>
- Otis, A. S., & Knollin, H. E. (1921). The reliability of the Binet scale and of pedagogical scales. *Journal of Educational Research*, 4(2), 121–142. <https://doi.org/10.1080/00220671.1921.10879187>
- Pearson, K. (1896). Mathematical contributions to the theory of evolution: III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London, Series A*, 187, 253–318. <https://doi.org/10.1098/rsta.1896.0007>
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5*, 50(302), 157–175. <https://doi.org/10.1080/14786440009463897>
- Pearson, K. (1904). On the laws of inheritance in man: II. On the inheritance of mental and moral characters in man, and its comparison with the inheritance of physical characters. *Biometrika*, 3(2–3), 131–190. <https://doi.org/10.1093/biomet/3.2-3.131>
- Pearson, K. (1907). *Mathematical contributions to the theory of evolution: XVI. On further methods of determining correlation*. Dulau & Co.

- Pearson, K. (1934). Reply from Professor Karl Pearson. In *Speeches delivered at a dinner held in University College, London in honour of Professor Karl Pearson, 23 April 1934* (pp. 19–24). Cambridge University Press.
- Pearson, K., & Moul, M. (1927). The mathematics of intelligence: I. The sampling errors in the theory of a generalized factor. *Biometrika*, 19(3–4), 246–291. <https://doi.org/10.1093/biomet/19.3-4.246>
- Perez, C. (2002). Different tests, same flaws: Examining the SAT I, SAT II, and ACT. *The Journal of College Admission*, 177, 20–25.
- Phelps, R. P. (2007). *Standardized testing primer*. Peter Lang.
- Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, 30(1), 39–56. <https://doi.org/10.1007/BF02289746>
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability: Vol. 4. Contributions to biology and problems of medicine* (pp. 321–333). University of California Press.
- Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 19(1), 49–57. <https://doi.org/10.1111/j.2044-8317.1966.tb00354.x>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. MESA Press; University of Chicago. (Original work published 1960)
- Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Harvard University Press.
- Remmers, H. H., Shock, N. W., & Kelly, E. L. (1927). An empirical study of the validity of the Spearman–Brown formula as applied to the Purdue rating scale. *Journal of Educational Psychology*, 18(3), 187–195. <https://doi.org/10.1037/h0072665>
- Rice, J. M. (1913). *Scientific management in education*. Publishers Printing Company.
- Rice, J. M. (1897a). The futility of the spelling grind: I. *Forum*, 23, 163–172.
- Rice, J. M. (1897b). The futility of the spelling grind: II. *Forum*, 23, 409–419.
- Robinson, D. (2005). Profiles in research: Howard Wainer. *Journal of Educational and Behavioral Statistics*, 30(4), 465–476. <https://doi.org/10.3102/10769986030004465>
- Robinson, D. H., & Wainer, H. (2006). Profiles in research: Arthur Jensen. *Journal of Educational and Behavioral Statistics*, 31(3), 327–352. <https://doi.org/10.3102/10769986031003327>
- Ruch, G. M., Ackerson, L., & Jackson, J. D. (1926). An empirical study of the Spearman–Brown formula as applied to educational test material. *Journal of Educational Psychology*, 17(5), 309–313. <https://doi.org/10.1037/h0075480>
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Psychometric Society.
- Samejima, F. (1972). *A general model for free-response data* (Psychometric Monograph No. 18). Psychometric Society.

- Samelson, F. (1990). Was early mental testing: (a) racist inspired, (b) objective science, (c) a technology for democracy, (d) the origin of the multiple-choice exams, (e) none of the above? (Mark the RIGHT answer). In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 113–127). Rutgers University Press.
- Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1999). *CATBOOK computerized adaptive testing: From inquiry to operation* (Study Note No. 99-02). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Shields, J., Hanser, L. M., & Campbell, J. P. (2001). A paradigm shift. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 19–28). Taylor & Francis.
- Sitgreaves, R. (1963). Book review: *Probabilistic models for some intelligence and attainment tests* by G. Rasch. *Psychometrika*, 28(2), 219–220. <https://doi.org/10.1007/BF02289619>
- Sokal, M. M. (1990). James McKeen Cattell and mental anthropometry: Nineteenth-century science and reform and the origins of psychological testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 21–45). Rutgers University Press.
- Spearman, C. (1904a). “General intelligence” objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–293. <https://doi.org/10.2307/1412107>
- Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 18, 161–169.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295. <https://doi.org/10.1111/j.2044-8295.1910.tb00206.x>
- Spearman, C. (1930). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography* (Vol. 1, pp. 299–333). Clark University Press.
- Staff, Psychology Branch, Office Air Surgeon. (1944). The aviation cadet qualifying examination of the Army Air Forces. *Psychological Bulletin*, 41(6), 385–394. <https://doi.org/10.1037/h0059734>
- Stanley, J. C. (1966). Rice as a pioneer educational researcher. *Journal of Educational Measurement*, 3(2), 135–139. <https://doi.org/10.1111/j.1745-3984.1966.tb00869.x>
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). Wiley.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Belknap Press.
- Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press.

- Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23(1), 57–75. <https://doi.org/10.2307/116534>
- Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of item in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Springer. https://doi.org/10.1007/0-306-47531-6_9
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2), 201–210. <https://doi.org/10.1177/014662168300700208>
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292. <https://doi.org/10.1177/014662169301700308>
- Stocking, M. L., & Swanson, D. (1998). Optimal design of item banks for computer adaptive testing. *Applied Psychological Measurement*, 22(3), 271–279. <https://doi.org/10.1177/01466216980223007>
- Stocking, M. L., Swanson, L., & Pearlman, M. (1991a). *Automatic item selection (AIS) methods in the ETS testing environment* (ETS Research Memorandum No. RM-91-05). ETS.
- Stocking, M. L., Swanson, L., & Pearlman, M. (1991b). *Automated item selection using item response theory* (ETS Research Report No. RR-91-09). ETS. <https://doi.org/10.1002/j.2333-8504.1991.tb01375.x>
- Stonehill, R. M., & Anderson, J. I. (1982). *An evaluation of ESEA Title I—program operations and educational effects. A report to Congress*. U.S. Department of Education, Budget, and Evaluation, Office of Planning.
- Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. L., Star, S. A., & Clausen, J. A. (1950). *Measurement and prediction*. Princeton University Press.
- Terman, L. M. (1917). The intelligence quotient of Francis Galton in childhood. *The American Journal of Psychology*, 28(2), 209–215. <https://doi.org/10.2307/1413721>
- Terman, L. M. (1922). The great conspiracy, or the impulse imperious of intelligence testers, psycho-analyzed and exposed by Mr. Lippmann. *The New Republic*, 33(421), 116–120.
- Terman, L. M., Baldwin, B. T., Bronson, E., DeVoss, J. C., Fuller, F., Goodenough, F. L., Kelley, T. L., Lima, M., Marshall, H., Moore, A. H., Raubenheimer, A. S., Ruch, G. M., Willoughby, R. L., Wyman, J. B., & Yates, D. H. (1925). *Mental and physical traits of a thousand gifted children: Vol. 1. Genetic studies of genius*. Stanford University Press.
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501–519. <https://doi.org/10.1007/BF02302588>
- Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation*, 1(1), 23–39. <https://doi.org/10.59863/GPML7603>

- Thorndike, E. L. (1903). *Educational psychology*. Lemcke & Buechner. <https://doi.org/10.59863/GPML7603>
- Thorndike, E. L. (1913). *Educational psychology, Vol. 1: The original nature of man*. Teacher's College, Columbia University.
- Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements*. Columbia University, Teachers College.
- Thorndike, R. L. (1947). *Research problems and techniques* (Vol. 3). Army Air Forces, U.S. Government Printing Office.
- Thorndike, R. L. (1949). *Personnel selection; Test and measurement techniques*. John Wiley.
- Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). American Council on Education.
- Thorndike, E. L., Lay, W., & Dean, P. R. (1909). The relation of accuracy in sensory discrimination to general intelligence. *American Journal of Psychology*, 20, 364–369.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433–451. <https://doi.org/10.1037/h0073357>
- Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology*, 18(8), 505–524. <https://doi.org/10.1037/h0075524>
- Thurstone, L. L. (1931). *The reliability and validity of tests*. Edwards Brothers.
- Tinker, M. A. (1932). Wundt's doctorate students and their theses 1875–1920. *The American Journal of Psychology*, 44(4), 630–637. <https://doi.org/10.2307/1414529>
- Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.
- Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1–13. <https://doi.org/10.1007/BF02288894>
- Urban, W. J. (2010). *More than science and Sputnik: The National Defense Education Act of 1958*. University of Alabama Press.
- van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Kluwer. https://doi.org/10.1007/0-306-47531-6_2
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. Springer. <https://doi.org/10.1007/978-1-4757-2691-6>
- Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1–22). Lawrence Erlbaum Associates.
- Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton University Press.
- Wainer, H., & Kiely, G. L. (1987). Item clusters in computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185–201. <https://doi.org/10.1111/j.1745-3984.1987.tb00274.x>
- Wainer, H., & Robinson, D. H. (2006). Profiles in research: R. Darrell Bock. *Journal of Educational and Behavioral Statistics*, 31(1), 101–122. <https://doi.org/10.3102/10769986031001101>

- Wainer, H., & Robinson, D. H. (2009). Profiles in research: Linda S. Gottfredson. *Journal of Educational and Behavioral Statistics*, 34(3), 395–427. <https://doi.org/10.3102/1076998609339366>
- Wallace, A. R. (1858). On the tendency of varieties to depart indefinitely from the original type. *Zoological Journal of the Linnean Society*, 3, 46–50.
- Webster, L. (1998). David Andrich: A genius from down under. *Popular Measurement*, 1, 26. <https://www.rasch.org/pm/pm1-26.pdf>
- Weinland, T. P. (1973). A history of the I.Q. in America, 1890–1941. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 33(10-A), 5666.
- Weiss, D.J. (Ed.). (1978). *Proceedings of the 1977 computerized adaptive testing conference*. University of Minnesota.
- Weiss, D.J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. Academic Press.
- Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements*, 3(6), i–62. <https://doi.org/10.1037/h0092995>
- Wolf, T. A. (1973). *Alfred Binet*. University of Chicago Press.
- Wood, B. D. (1926). Studies in achievement tests. Part III. Spearman–Brown reliability predictions. *Journal of Educational Psychology*, 17(4), 263–269. <https://doi.org/10.1037/h0063759>
- Wood, D. A. (1962). *Louis Leon Thurstone: Creative thinker, dedicated teacher, eminent psychologist*. ETS.
- Wright, B. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 ETS Invitational Conference on Testing Problems* (pp. 85–101). ETS.
- Wright, B. (1992, April 20–24). *IRT in the 1990s: Which models work best? 3PL or Rasch?* [Paper presentation]. American Educational Research Association Annual Meeting. San Francisco, CA, United States. <https://www.rasch.org/rmt/rmt61a.htm>
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). American Council on Education; Praeger.
- Yerkes, R. M. (Ed.). (1921). *Psychological examining in the army* (Vol. XV). National Academy of Sciences.
- Yoakum, C. S., & Yerkes, R. M. (Eds.). (1920). *Mental tests in the American army*. Sidgwick & Jackson. <https://doi.org/10.1037/11054-000>
- Yule, G. U. (1897). On the significance of Bravais' formulæ for regression, &c., in the case of skew correlation. *Proceedings of the Royal Society of London*, 60, 359–367. <https://doi.org/10.1098/rspl.1896.0075>
- Zenderland, L. (1990). The debate over diagnosis: Henry Herbert Goddard and the medical acceptance of intelligence testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 46–74). Rutgers University Press.

- Zenderland, L. (1998). *Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing*. Cambridge University Press.
- Zhang, H., & Luo, F. (2020). The development of psychological and educational measurement in China. *Chinese/English Journal of Educational Measurement and Evaluation*, 1, 56–64. <https://doi.org/10.59863/BUAI8988>

NOTES

1. The first and second wranglers that same year are decidedly less well known: Andrew James Campbell Allen and George Francis Walker. Other well-known scholars who performed well in the examinations but failed to achieve first place include Alfred North Whitehead, John Venn, Bertrand Russell, Thomas Malthus, John Maynard Keynes, and R. A. Fisher.
2. Several other formulations of this equation may be more familiar to the reader, including $r_{xy} = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2 \Sigma y^2}}$.
3. This statement obviously reflects Pearson's enthusiasm for eugenics, but it should be noted that it was made years before the atrocities of the Holocaust were widely known.
4. The form of the equation provided in the text is the one presented by Spearman. The reader may be more familiar with the following equivalent formulation in which x and y replace p and q , $r_{x'y'} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$.
5. The reader may be familiar with the following formulation in which r_{11} is the reliability of the original test and r_{KK} is the reliability of a test lengthened by a factor of K , $r_{KK} = \frac{K r_{11}}{1 + (K - 1) r_{11}}$.
6. Clauser (2021) provided a more detailed presentation of the material in this section.
7. These papers were originally published in 1905, 1908, and 1911. In 1916, at Goddard's instigation, they were translated and published in a single volume. In what follows, the chronology of development will be based on the original publication dates, but specific references will be to the translation.
8. Some previous authors explicitly state that Binet and Simon may have been the first psychologists to work with item characteristic curves (e.g., Hambleton & Swaminathan, 1985). This requires a broad definition of *item characteristic curve*. Binet and Simon examined the probability of a correct response as a function of age rather than as a function of estimated ability. In this context, it is also worth noting that it does not appear that Binet and Simon actually plotted their results; instead, they presented the results in tabular form. That said, several of the tables in

their paper on application of the new methods present scores as a function of age and could scarcely be more informative if they were converted to graphic form.

9. Our copy was signed by Goddard in 1949 with a note supporting such beneficial outcomes claiming that “Deborah Kallikak still lives, not in story but in fact. She is now 58 years old in good health and enjoying life . . . across the street from the Vineland Training School where she was living when this book was written.”
10. The idea of forming a quotient by dividing mental age by chronological age was originally recommended by the German psychologist William Stern, but the widespread use and its transfer to a three-digit scale—with 100 representing a 1:1 ratio of mental age to chronological age—is the result of Terman’s work.
11. Not to be mistaken for the publishers of the *World Book Encyclopedia*.
12. The Terman study may be the longest running psychological study. The Fels study ran longer—from 1929 to 2018—however, this study was primarily focused on physiological measurements.
13. In addition to Terman’s effort to follow individuals into adulthood to examine how their exceptional IQs impacted their lives, he apparently thought it was possible to estimate the IQs of historical figures; he published a paper in which he estimated Francis Galton’s childhood IQ by evaluating his childhood mental age based on records of his childhood behavior. He concluded that his IQ must have been around 200 (Terman, 1917).
14. Brigham subsequently changed his mind about the usefulness of intelligence tests for making racial comparisons. In Brigham (1930), he made this change of opinion clear, stating in reference to *A Study of American Intelligence*, “This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies—the writer’s own—was without foundation” (p. 165).
15. Even Helen Keller supported the eugenics movement. In 1915, she wrote a letter to *New Republic* supporting a doctor who withheld treatment from a severely disabled baby. In that letter she advocated the formation of juries of physicians to make such decisions, stating, “A mental defective . . . is almost sure to be a potential criminal. The evidence before a jury of physicians considering the case of an idiot would be exact and scientific. . . . They would act only in cases of true idiocy, where there could be no hope of mental development.” She went on to say, “Conservatives ask too much perfection of these new methods and institutions, although they know how far the old ones have fallen short of what they were expected to accomplish. We can only wait and hope for better results as the average of human intelligence, trustworthiness and justice arises.”
16. Readers may be more familiar with this formulation in which the subscript t

signifies true score and the subscript x signifies total score: $r_{12} = \frac{\sigma_t^2}{\sigma_x^2}$.

17. It is worth noting that a paper by Cyril Hoyt (1941) published in *Psychometrika* a decade earlier (and cited by Cronbach) also demonstrated that the KR20 represented the mean of all possible split-half estimates. Hoyt's paper also provided a link to generalizability theory by producing the estimate based on analysis of variance.
18. The suggestion that error of measurement might be a more useful metric than a reliability coefficient did not originate with Cronbach. Otis and Knollin (1921) and Kelley (1921) both made this argument. As we noted, the histories of the classical test theory and generalizability theory perspectives are intertwined.
19. This is a simplified description of generalizability theory. For a more complete description, see Cronbach et al. (1972) or Brennan (2001a). Brennan (2021) specifically presented aspects of the history of generalizability theory.
20. In addition to this, Edgeworth also published two papers on correlation years before Pearson's paper and he coined the term *coefficient of correlation*.
21. In the second edition of *Educational Measurement*, William Angoff wrote a chapter titled "Scales, Norms, and Equivalent Scores" (Angoff, 1971). Angoff viewed scaling as the process of defining the scale on which scores are to be reported.
22. This relationship would have been familiar to Fechner through his research on electricity. Georg Ohm had written that the loss of current would be a logarithmic function of the length of the wire: $V = m \text{ Log } (1 + x)$.
23. Lord also proposed the less well-known four-parameter IRT model.
24. In this context, it is worth noting that Lord was substantially focused on models that were applicable to multiple-choice questions. Rasch's interests were more general. Rasch (1960/1980) presented a model for reading accuracy, a model for reading speed, and a model for intelligence tests. In reviewing Rasch's book for *Psychometrika*, Sitgreaves (1963) noted that the first two of these models are more fully developed than the third.
25. Wright (1968) was based on a paper presented at ETS in 1967, but Wright did not actually discuss the model in that paper—it is included as a footnote, which may have been added later.
26. Thissen and Steinberg (2020) attributed the use of the term one-parameter model to Wainer.
27. The research reported in this volume began with his study of spelling in 1895. Work on mathematics and language followed in 1901 and 1903, respectively.
28. Kelley (1927) considered this an inappropriate practice. He commented, "On the average, in the neighborhood of .90 of the capacity measured by an all-round achievement battery score,—reading arithmetic, science, history, etc.—and the capacity measured by a general intelligence test is one and the same. If a comprehensive educational achievement test and a general intelligence test each give 'fairly reliable' total scores, each would need to be more than ten times as

long to yield equally reliable measures of the difference between the educational achievement and intelligence scores” (pp. 21–22). It may be that the intentions behind these approaches to testing differ more than the results.

29. SAT was originally an initialism representing Scholastic Aptitude Test. The name was subsequently changed to Scholastic Assessment Test and is now officially SAT.
30. Similar to the SAT, ACT originally was an initialism for American College Testing. ACT is now the official name of the test.
31. For more information on the evolution of NAEP, see Jones and Olkin (2004).
32. This information is inferred from the names of the authors. In the case of K. W. Vaughn, this inference was not possible and we were unable to find other specific information about this author.

The Sociocultural Context of Educational Assessment

Kadriye Ercikan

ETS

Guillermo Solano-Flores

Stanford University

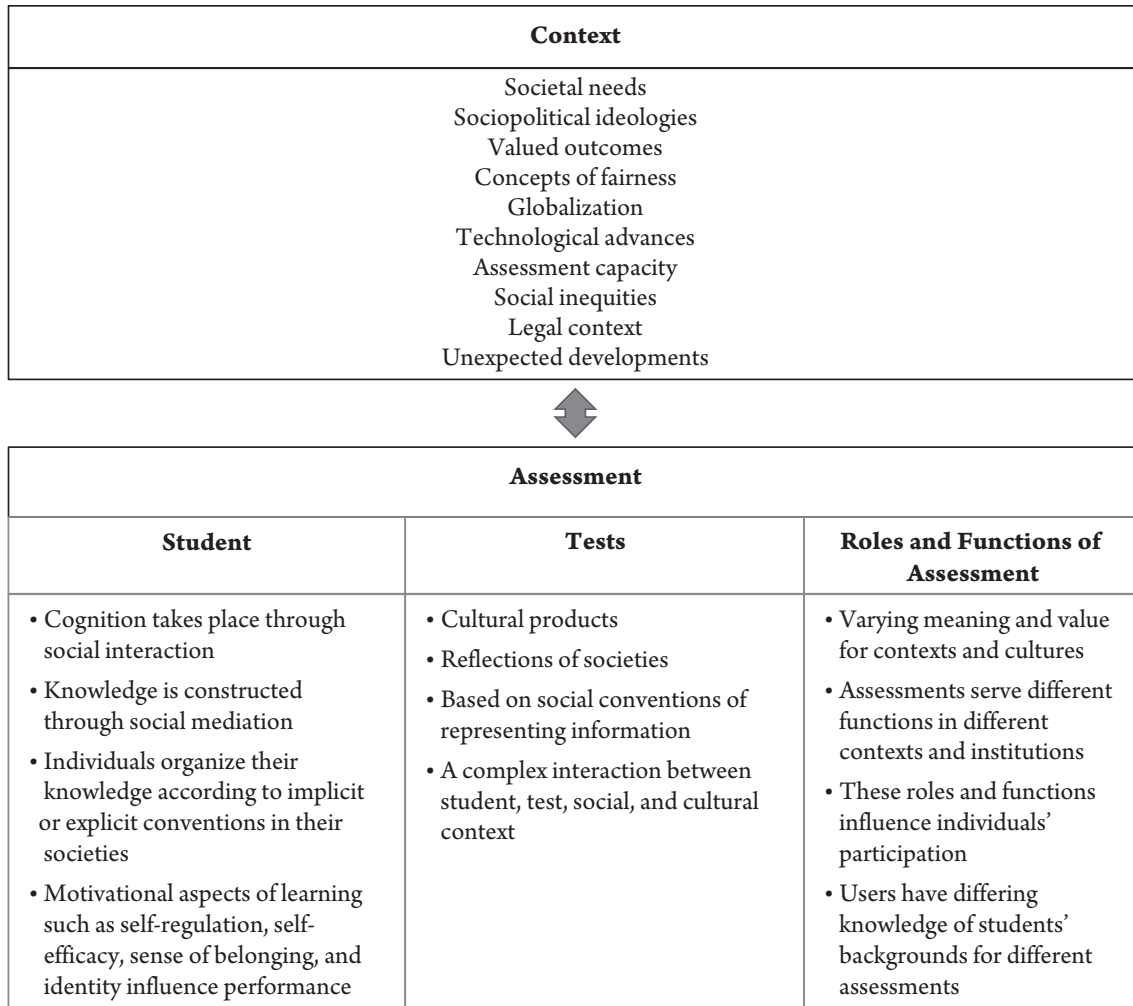
In this chapter, we argue for and demonstrate the critical importance of considering assessment as an activity that both reflects and influences the sociocultural contexts in which it is developed and enacted. Our main goal is to highlight the nature of this bidirectional relationship for measurement professionals, researchers, decision-makers, and practitioners. We provide guidance for research and practice that contribute to promoting valid and equitable design, interpretation, and uses of assessment while taking this bidirectional relationship into account. Contexts determined by conditions such as legislation, educational policy and practices, technological advances, social inequities, and unexpected events such as the COVID-19 pandemic affect assessment practices and individuals' participation in assessment (e.g., Tong, 2022). In turn, assessment impacts education policy, teaching practices, and the learning, career, and life opportunities of individuals. There has been a great deal of attention paid to this bidirectional complex relationship recently by scholars emphasizing different aspects of the relationship (Bennett, 2023; Hughes, 2023; Randall, 2021; Russell, 2023; Walker et al., 2023). One of the key goals of our chapter is to promote dialog among stakeholders with varying perspectives on social justice and sociocultural issues given the multiplicity of stances and real-world goals of stakeholders.

This sociocultural perspective on the complex bidirectional relationship between assessment and the sociocultural context is summarized in Figure 3.1. The figure is far from being exhaustive and is intended to illustrate the variety of elements involved in the relationship. In representing this relationship, we embrace a sociocultural perspective (Wertsch, 1985) according to which cognition takes place through social interaction (National Academies of Sciences, Engineering, and Medicine [NASEM], 2018b; Rogoff & Lave, 1984; Vygotsky, 1978): Knowledge is constructed through social mediation; individuals make sense of experience according to implicit or explicit conventions about what is accepted or deemed reasonable in their societies. Different epistemologies result from different cultural influences that shape how individuals make sense of experience and how they view and solve problems (Greenfield, 1997).

We also pay attention to language and its associated communicative roles in institutional and social life as both a cultural product and a key mediator of the ways in which knowledge is represented and information is understood (Gelman & Roberts, 2017). Because languages encode cultural experience accumulated over time, each language has a unique set of resources and conventions that shape the way in which ideas are expressed.

Three notions derived from this perspective guide our reasoning about the bidirectional relationship of assessment and sociocultural contexts (Ercikan & Solano-Flores, 2016). The first notion is that culture and language influence how individuals engage with tests, how they process information presented in tests, and how they formulate solutions and generate responses.

The second notion is that, because tests and testing are cultural products (Cole, 1999), they are a reflection of the societies in which they originate and are even an agent of social reproduction. The content of tests is a reflection of the kinds of knowledge, reasoning,

**FIGURE 3.1**

The Bidirectional Relationship Between Sociocultural Context and Assessment

and skills valued in a society (Hughes, 2023; Teltemann & Klieme, 2016). Also, the ways in which tests are worded and the ways in which they represent information are aligned with implicit and explicit social conventions for representing information based on multiple assumptions about what is understandable to the test taker (Solano-Flores et al., 2013). Tests tend to be written in a standard, formal variety of the language in which they are administered and are aligned with immediate performance of tasks and their demands presented in artificial events. This type of language usage is not common among certain test takers—and especially so for test takers from a nonnative language background in the language of a test (Solano-Flores, 2006).

The third notion is that assessment has different meaning and roles for different cultures (e.g., Gebril, 2016; Kennedy, 2016) and contexts. Views of tests and testing and the consequences that tests and performance on tests have on people's lives vary across

societies. This creates differences across societies in the ways in which individuals prepare for taking tests, how they interpret test items and their demands, the level of effort they put into responding to tests, and the pace with which they solve problems during a test.

Even though much of the discussion is relevant to both large-scale assessment and classroom assessment, we focus intentionally only on large-scale assessment, given the importance of addressing the wider range of stakeholders, contexts, and formalized processes involved in the bidirectional relationship between sociocultural context and assessment. Discussions of classroom assessment and its relationship with large-scale assessment systems can be found in Ruiz-Primo and Furtak (2024) and in Brookhart and DePascale (this volume).

LIMITATIONS IN CURRENT PRACTICES

Despite recognition of the complex relationships between assessment and society, the nature of social contexts may not always be given sufficient consideration in current large-scale assessment practices. In this section, we discuss challenges and limitations in four areas of research and practice: assessment interpretation and use, assessment design and development, the operationalization of culture and language, and approaches to fairness research.

Assessment Interpretation and Use

A key implication of the sociocultural perspective is that test score interpretation needs to be grounded in the sociocultural contexts of the participating individuals and the assessment itself. For example, differences in performance on tests largely reflect social and cultural group differences that are confounded with the actual targeted constructs. In some assessments, the targeted construct is identified in vague terms such as *verbal ability*, which implies a more general reading ability than what is targeted by the assessment. For example, the assessment may in fact target reading ability to carry out a set of reasoning tasks in specific contexts that involve text. Since the context of the application of reasoning is not specific, the assessed reading ability may be perceived as “general, abstract reading ability.” The exclusion of the context from the construct definition obscures the critical need for inclusion of the context for applying the targeted skills and abilities. It also masks the need for the type of evidence required for the intended interpretation and use of the assessment, which is likely to interact with sociocultural factors. A wealth of literature and data document the considerable score gaps on large-scale assessments between groups defined by factors such as socioeconomic status, race/ethnicity, disability, and proficiency in the language of tests (e.g., Bohrnstedt et al., 2015; Gilmour et al., 2019; NASEM, 2018a). In the United States, compared to their middle-class and high-income counterparts, low-income students, Black and Latino students, and English learners (ELs; students whose first language is not the language in which they receive instruction and are tested) attend schools with fewer resources, tend to be taught by less qualified teachers, and are less likely to be

exposed to a curriculum that contains all the content assessed (Darling-Hammond, 2000; NASEM, 2017). While assessment programs are intended to serve the goal of producing insights about educational outcomes, they cannot inform policy and practice effectively without taking these important aspects of sociocultural contexts into account.

Recognizing that sociocultural contexts shape the validity and interpretation of test results is not new (see Becker, 2003; Ercikan, 1998; Ercikan & Roth, 2011). However, the focus has been primarily on examining the extent to which tests and the scores they produce can be assumed to have the same meaning when they are administered in different versions or to individuals from different groups (see Berman et al., 2020). In these comparisons, psychometric approaches such as the analysis of differential item functioning (DIF; Camilli, 2013; Hambleton et al., 2005) are frequently invoked as evidence of the ability of assessment systems to ensure fairness by identifying test items that are biased against a given group. Yet a critical perspective of empirical evidence reveals that such approaches are not sensitive to cultural, linguistic, and economic heterogeneity *within* each of the groups compared, despite evidence that cultural and linguistic heterogeneity is an important source of bias and measurement error (see Ercikan & Por, 2020; Oliveri et al., 2013; Solano-Flores & Li, 2009).

The need for critically examining the influence of sociocultural contexts on student performance is also evident in our efforts to examine response processes. The analysis of cognitive activity during test taking has been part of the arsenal of techniques for test development and test validation since the early 1980s (e.g., Chi et al., 1981; Hamilton et al., 1997; Kukea Shultz & Englert, 2021; Leighton, 2013; Pepper et al., 2018; Ruiz-Primo et al., 2001). Yet rarely has examination of evidence of cognitive activity taken into consideration the social nature of knowing and learning or the role that the sociocultural context plays in learning (see Gee, 2008; National Research Council [NRC], 2001). A sociocultural perspective should allow examination of student performance on tests in terms of how individuals think, feel, act, and interact (see Moss et al., 2008; Nasir, de Royston, et al., 2020; NRC, 2000; Solano-Flores & Nelson-Barber, 2001). Cultural factors, threats to validity due to limited language proficiency in the language of tests, and differential predictive validity of tests across individuals from different cultural groups (see van de Vijver, 2016) occupy a prominent role among the factors that deserve our attention.

Assessment Design and Development

In practice, past approaches to addressing culture and language in assessment development have largely focused on the presentation-related features of items that are known to create unnecessary linguistic complexity (e.g., Abedi, 2006) and on the inclusion of assessment developers and reviewers who represent the student target population and who are assumed to be sensitive to its characteristics and assessment needs (Welch, 2006; Zieky, 2016). Although they are valuable, these approaches may be limited in their ability to address the complex influences of the sociocultural context and language in testing. As mentioned, inevitably, tests are created not only based on

explicit prescriptions of the content that students should know but also under implicit assumptions about both the shared cultural experiences that make tests accessible and the students' familiarity with the response demands of tests as they are administered. Often, as a result, the linguistic features of tests, the contextual information they provide with the intent to make items meaningful, and the forms in which they represent information may reflect the experience and culture of those who create the tests more than the test takers' experiences. These concerns are particularly serious given that there is evidence that students rely on out-of-school, firsthand, everyday life experiences to make sense of the content of items and the contextual information they provide (Hamilton et al., 1997; Solano-Flores & Li, 2009).

Currently accepted practice in assessment design and development is not guided by clear criteria to systematically address the ways of knowing of different populations and sociocultural contexts (see Faulkner-Bond, 2020; Randall, 2021). Researchers have expressed concerns about how constructs are represented and the fact that the wording of test items and the contextual information they provide exclude certain segments of the student populations (e.g., Durán, 2011; Randall, 2021). However, very scant research (e.g., Banks, 2006; Martiniello, 2008) has examined the characteristics of test items according to variables relevant to test takers who are familiar with relevant cultural contexts and studied how those characteristics are associated with differences in student performance. Recent evidence on the development of assessments for Indigenous populations in Hawaii (Kukea Shultz & Englert, 2021) indicates that working intensively and directly with members of local communities allows development of meaningful test items. While this kind of work may not necessarily be feasible for large-scale assessment programs and heterogeneous populations, this evidence speaks to the importance of including samples of students from different cultural, linguistic, and socioeconomic groups in the process of test development as critical to ensuring fair assessment (for example, including them in cognitive interviews and focal groups intended to contribute information that allows examination of inferred cognitive processes elicited by pilot items).

The exclusion or limited inclusion of ELs in the process of assessment development is a special concern in the United States because ELs constitute the population that is most likely to be affected by challenges imposed by the linguistic features of test items. While these students are part of the population tested in large-scale assessment programs, in which tests are administered in English, they are typically not adequately represented in pilot studies to determine how the wording of items may pose challenges for understanding the items as intended and responding to them. Underlying this practice may be the erroneous conception that ELs do not have the communication skills needed to participate in responding to draft versions or participating in cognitive interviews or focus groups. In actuality, despite their classification as ELs, the majority of these students have at least a minimum level of English proficiency and ought to participate in the pilot stages of the development of tests to provide valuable information on the ways in which the wording and format of items can affect their ability to understand them (see Kachchaf, 2011; Winter et al., 2006; Young et al., 2014).

Evidence of cultural relevance and appropriateness of assessment tasks for students from different cultural backgrounds and identities is even rarer and more limited (Randall, 2023). A key, simple way of examining cultural and linguistic appropriateness of items consists of conducting cognitive labs with individuals from different backgrounds to probe their interpretations of test items and the variety of ways in which they respond to them. Unfortunately, in practice, these activities do not involve adequate samples of students from different cultural, linguistic, and socioeconomic groups. At best, broad categories of linguistic or ethnic groups are used under the implicit assumption of population homogeneity—a serious limitation, given the fact that cultural and linguistic minority groups tend to be highly heterogeneous (Ercikan & Roth, 2009; Sireci & Faulkner-Bond, 2015). Missing this opportunity for examining the degree to which test items align with students' cultural experiences and language can compromise measurement and the interpretation of assessment results.

Operationalization of Culture and Language

An important challenge in addressing cultural and linguistic issues in assessment is the difficulty of developing actionable principles that are sensitive to the characteristics of the populations included. There are comprehensive guidelines for developing tests for diverse populations offered by both scholars (e.g., Taylor, 2022) and professional organizations (see American Educational Research Association [AERA] et al., 2014; International Test Commission [ITC], 2017). In addition, most assessment programs have item-writing guidelines intended to address culture and language (e.g., ETS, 2022). Large-scale assessment programs include stages in the process of assessment development in which external reviewers examine items to identify potential sources of bias or difficulty in the wording of the items or the contexts they use.

Unfortunately, while these actions may produce valuable information on how the characteristics of items need to be adjusted, the information provided by adult reviewers may not be comparable to the information obtained from students with whom the items could be tried out. There is evidence that the features of items reviewers identify as potentially challenging for proper interpretation by their students are not necessarily the same features that pose a challenge to those students (Solano-Flores & Trumbull, 2003). The absence of proper conceptual and ideological frameworks imposes a limit to the ability of test reviewers to detect and address limitations concerning issues of culture and race (Randall, 2023).

Regarding the testing of ELs, for example, it is not uncommon to find recommendations for item writers that focus on principles for the use of linguistic features such as word length, negation, conditional clauses, and passive voice (e.g., Abedi, 2006, 2016). While these recommendations may be well informed by empirical research, they are limited in their ability to minimize unnecessary linguistic complexity because they are not item specific. Analyses of test takers' interpretations of items always reveal unanticipated challenges. Indeed, it is extremely difficult to predict what may become an issue in the development of an item. Not only does each item have a unique combination of

linguistic and contextual features (in addition to the content it is intended to assess), but also each test taker has a unique combination of skills and cultural experiences that influence how they make sense of items. Because of this variation and its unpredictability, actions oriented to addressing the cultural and linguistic aspects of tests need to be taken for each item and need to be empirically supported.

Limitations in addressing cultural diversity largely stem from the predominance of implicit deterministic views underlying assessment practice and policy and the role of schools in educating students with limited understanding of their backgrounds. It is a surprising fact that, while probability is at the core of the reasoning in the field of psychometrics, assessment systems' procedures reflect deterministic, rather than probabilistic, views of culture and language (Solano-Flores, 2014). At the core of these deterministic views is the treatment of populations as homogeneous. An example of these deterministic views can be seen in approaches for examining bias based on the analysis of DIF. These approaches implicitly assume homogeneity in both the focal and the reference samples, which may lead to a high proportion of undetected biased items (Ercikan et al., 2014). A second example of this deterministic view is the use of small numbers of general ordered categories to characterize EL students' proficiency in English. The limited number of categories and their oversimplification fail to communicate the complexity of a given student's reading, writing, and conversation skills in an EL student's second language and the interaction of these skills with each other in authentic situated communication (see Hsieh et al., 2023; Linquanti et al., 2016). A third example is observed in translation procedures used in international large-scale assessment comparisons, which are silent about linguistic variation within the users of the same language and within the same given country (e.g., Yu & Ebbs, 2012).

The limitations discussed point to the complexity and uncertainty involved in defining and operationalizing culture, language, disability, and any other form of diversity into a narrow set of categories. They also speak to the need for treating tests as boundary objects. Star and Griesemer (1989) proposed the notion of *boundary objects* in identifying common elements in complex interactions. This notion is defined as "an analytic concept of those scientific objects which both inhabit several intersecting social worlds . . . and satisfy the informational requirements of each of them" (p. 393). The notion of boundary objects is useful in conceptualizing aspects of tests that are flexible enough to adapt to local individual student backgrounds and needs, yet have elements that are common enough across students and contexts to support consistent meaning. The balance of commonality and uniqueness in tests to account for the complex interaction between student, context, and assessment is central to supporting consistency of meaning and claims.

FAIRNESS RESEARCH

Differential item functioning (DIF) methodology has constituted a key approach to examining fairness in assessments. Sociocultural perspectives challenge the assumption of within-group homogeneity in the statistical analyses used for detecting DIF. This is

especially important in the case of second-language learners, ethnically diverse groups, and students with disabilities—groups that can be tremendously heterogeneous. There is evidence that as the heterogeneity of focal group samples on sociodemographic variables increases, the detection of biased items decreases (Oliveri et al., 2014). This finding indicates that the violation of the homogeneity assumption has significant implications for fairness and bias analyses (Oliveri et al., 2014). In addition, DIF-based approaches commonly assume that measurement is comparable for the considered groups, *except* for the item being analyzed, given the reliance on performance on the assessment as the conditioning variable for comparisons. This assumption ignores the possibility that, given a set of intended interpretations and uses, a test as a whole is biased against a particular group.

There are also important limitations, which are often ignored, to the extent to which DIF-based approaches may mitigate bias. A first set of limitations stems from the fact that, while they allow detection of biased items, DIF-based approaches cannot detect the specific characteristics of items that may lead to bias. Work based on interviewing or conducting think-aloud protocols with students to elicit their interpretations of translated items has shown the complexity of disentangling sources of DIF (Ercikan, 2002; Kachchaf et al., 2016; Zenisky et al., 2004). Unless they are used in combination with cognitive interviews, verbal protocols, and other methods for probing cognitive activity, DIF-based approaches do not provide actionable directions to adjust the characteristics of differentially functioning items.

IMPLICATIONS OF A SOCIOCULTURAL PERSPECTIVE ON MEASUREMENT PRINCIPLES AND STANDARDS

From a sociocultural perspective, testing can be viewed as a multidirectional nexus among institutions, people, and purposes—all concerned with gathering, transmitting, and using information about what people know and can do. From this perspective, performance on a test can be viewed as the outcome of a complex set of interactions between an individual's knowledge and skills, personal and social background, learning experiences, and the conditions under which testing takes place. There are an infinite number of ways in which this complex interaction between sociocultural contexts and assessment can take place given the large numbers of students, stakeholders, contexts, and assessment characteristics and perspectives that might be at play. Differences in the meaning of this complex interaction create challenges to supporting claims and interpretations of assessment results that are consistent across individuals and groups, which requires finding common elements of the interaction that can support such consistency.

The field of measurement has been concerned about the properties of tests and how they are interpreted and used for different groups. Guidelines have been developed to “provide criteria for the development and evaluation of tests and testing practices and . . . the interpretation of test scores for the intended uses” (AERA et al., 2014, p. 1).

Standards such as those generated by AERA et al. (2014) and by the ITC (2017) acknowledge the importance of examining the properties of tests and the evidence that supports their interpretation and use for individuals from different groups focusing on broad demographic groups defined by race/ethnicity, gender, language, disability, etc. Central to the measurement principles communicated by these guidelines is the assumption of standardization—ensuring

that the results of measurement (the numeric values) will have an interpretation that is trustworthy and invariant (within specified limits) to the objects being measured, the instruments used to produce the numeric values, and the person who is interpreting the values. (Briggs, 2022, p. 13)

Efforts to standardize the conditions under which test takers are tested have a long history in educational measurement. Traditionally, standardization has been considered central to ensuring score comparability and fairness. It assumes that, even though some aspects of assessment may advantage some students, other aspects advantage other students, and on average, advantage or disadvantage evens out—which is referred to as *marginal inference* (Mislevy, 2018).

A sociocultural perspective challenges the notion that standardization and equivalent conditions and procedures result in uniformly *equivalent evidence* of students' knowledge, skills, and abilities in tests. There is ample evidence that standardization and equivalent conditions and procedures do not always provide *equivalent evidence* of students' knowledge, skills, and abilities. This is particularly true for minoritized student populations such as students who are schooled in a second language (Ercikan et al., 2014) and students with disabilities (Kopriva et al., 2016), or across student populations from different cultural backgrounds within countries or across countries (Ercikan et al., 2015).

As an alternative to marginal inference, *conditional inference* departs from observing everyone under the same conditions to a situation in which everyone's performance is evaluated by procedures targeting the same construct (Mislevy, 2018). Legislation requiring the inclusion of all students in large-scale tests has led to questioning whether equal testing administration conditions suffice to ensure valid, fair assessment for diverse populations (DePascale & Gong, 2020). Testing accommodations (adjustments made on the conditions on which tests are administered) are an example of a move toward flexibility in the ways in which test items are presented, the ways in which students can respond to those items, the settings used to test students, and the timing of tests, among other things (Kopriva et al., 2016; Mislevy et al., 2013; Thurlow & Kopriva, 2015). For example, a read-aloud option or assessment in a different language deviates from a standardized version with the intent to allow students to access, interact with, and provide responses to tasks in modes that are aligned with their backgrounds (Rose et al., 2012).

Unsurprisingly, validity is at the core of this tension between standardization and flexibility. On the one hand, standardization may restrict the opportunities for atypical or minoritized students to demonstrate knowledge because of limited language

proficiency, disability, cultural differences, etc. On the other hand, accommodating a wide spectrum of students' needs may be a formidable and costly task complicated by the lack of appropriate theories on testing accommodations, uncertainty about the fidelity with which accommodations are implemented, and lack of sufficient empirical evidence of the effectiveness of multiple forms of testing accommodations used by different assessment programs.

Treating observations as equivalent across students from different sociocultural backgrounds has implications for all aspects of measurement, including validity of score interpretations, assessment design and development, measurement modeling, and concepts of fairness. The view of tests as boundary objects highlights the possibility of designing tests that are flexible enough to adapt to local individual student backgrounds and needs with respect to content, presentation, and format and yet have elements that are common enough to facilitate consistent meaning, interpretation, and use (Bennett, 2023). However, there are multiple conceptual challenges that need to be addressed before such an ideal can be achieved. Among other things, existing testing practices (e.g., assessment development, design, and review) need to be revisited and important steps need to be taken to ensure that cultural, linguistic, and ability diversity are addressed and enforced with more rigor by professional organizations and testing standards (Solano-Flores, 2023).

The notion of conditional inference opens the door for developing interpretation and measurement models that allow creation of equivalent evidence of the construct from observations based on a combination of flexible and common elements of tests and facilitate equitable and consistent interpretation and use of tests. The challenges posed by a sociocultural perspective to the field of measurement are foundational and significant. This is a critical time for the field to recognize the consequences of ignoring this perspective on equity and to rectify this situation as a fundamental validity issue. In the following sections, we first discuss assessment as a social phenomenon and elaborate on the bidirectional relationship between sociocultural context and assessment. We then describe the bidirectional relation of legal context and assessment as an example of this strong bidirectional relationship. This is followed by a section on designing tests that are better grounded in their underlying sociocultural contexts using an evidence-centered design approach. We end the chapter with a discussion of the implications of sociocultural contexts on validity and fairness of interpretation and use.

ASSESSMENT AS A SOCIAL PHENOMENON

In the social and behavioral sciences, measurement is always a challenge, not only because many of the constructs of interest are latent (not visible), but also because each measurement endeavor involves a unique set of interacting factors (see Cronbach, 1975; Hedges, 1987). It is not an overstatement to say that, in educational measurement, many of those interacting factors are sociocultural in nature. Indeed, psychological attributes “are dynamically indexed to particular cultural, social, and historical conditions”

(Maul et al., 2016, p. 312). Obvious examples of this social, cultural, and historical relativism are the challenges encountered when attempts have been made to measure the quality of family functioning in international contexts (Puffer et al., 2021). What counts as “quality” or “functioning” varies tremendously across cultures. Therefore, how effectively the construct is operationalized based on the relationships between family members may be limited by these social, cultural, and historical contexts.

The complex relationship between assessment and society is largely due not only to the challenges of measuring psychological and social constructs—an issue that has been discussed for a long time in psychometrics literature (e.g., Green, 1981)—but also to the challenges of understanding educational measurement. Unlike measurement practices in other professions (e.g., medicine, engineering), measurement in education is particularly likely to affect the constructs being measured. For example, there is evidence and concern that what is tested tends to influence what is taught (Jenning & Bearak, 2014; Popham, 2001). Moreover, probably more than in any other discipline and profession, measurement in education is regulated or at least influenced by societies in ways that determine what is to be assessed and how constructs are to be measured. Furthermore, unlike measurement and measures relevant to other aspects of life (e.g., inflation, pollution, nutritional content, cholesterol levels), measurement in education is the one in which almost everyone has a stake, has some form of first-hand experience with, and believes to know well enough to have an authorized opinion about.

This section provides a model that represents the complex relationship of multiple aspects of assessment and multiple aspects of a society, followed by a discussion of different roles assessment plays in this relationship.

Sociocultural Influences

While it is widely recognized that social interaction provides individuals with opportunities to learn through collaboration and dialogue (NRC, 2000; Tomasello, 2009; Vygotsky, 1978), the notion is much more than a good pedagogical principle. Evidence from multiple disciplinary perspectives, including sociocultural theory, learning sciences, and neuroscience, speaks to the influence of culture and society on learning (e.g., Cole, 1999; Cole et al., 1971; Kahneman, 2003; Meltzoff et al., 2009; NASEM, 2018b; Nasir, Lee, et al., 2020; NRC, 1999, 2001; Rogoff et al., 2003; Roth & Lee, 2007). This perspective differs from cognitive views of learning focused mainly on “the way knowledge is represented, organized, and processed in the mind” (NRC, 2001, p. 3), mostly silent about society and culture.

A sociocultural perspective on learning holds that culture and society shape the mind (Wertsch, 1985). According to this sociocultural perspective, learning takes place through both informal everyday life and formal schooling experiences. The former takes place through the individual’s participation in communities of practice and plays a critical role in shaping individuals’ worldviews, identities, and the ways in which they make sense of different kinds of knowledge (Greeno, 2002; Greeno & Gresalfi, 2008;

Lave & Wenger, 1991; NASEM, 2018b). The latter are shaped, among many other factors, by structured activities, curriculum, resources, and teacher practices (Goldman & Lee, 2024; Greeno & Gresalfi, 2008; Moss et al., 2008). Altogether, these experiences both reflect and influence how individuals organize knowledge and solve problems, what kinds of knowledge are valued, and what counts as learning outcomes in a society.

Considering how culture and society shape the mind, we need also to examine the impact of social stressors on learning. For example, in the United States, individuals who self-identify as Black, Latino, Native American, gay, or transgender are more likely than mainstream individuals to have a lower quality of life, as shown by most metrics of well-being, such as physical and mental health, employment, housing, education, income, and wealth (Verschelden, 2017). Also, minoritized populations are more likely to suffer chronic illnesses such as asthma, diabetes, and heart disease. There is evidence that these chronic diseases are associated with higher school absenteeism and psychological problems among children (Chavira et al., 2004; Martinez & Ercikan, 2009; Peterson et al., 2006; Shaw & Páez, 2002).

Social stressors such as poverty, bullying, and violence are well known to lead to social withdrawal, peer conflicts, and difficulty developing adaptive behavior and healthy peer relationships (Shaw & Páez, 2002; Taras & Potts-Datema, 2005). These stressors also lead to an increased likelihood of developing anxiety and anxiety-related disorders (Adewuya & Ola, 2005; Baki et al., 2004; Chavira et al., 2004; Hommel et al., 2003)—factors that are shown by extensive research to be associated with lower performance on achievement tests and an increased incidence of learning difficulties and disabilities (Albano et al., 2003; Huberty, 2012; Leibert & Morris, 1967; Shaw & Páez, 2002; Taras & Potts-Datema, 2005; Van Ameringen et al., 2003; Wigfield & Eccles, 1990; Wine, 1971, 1980).

Research on the sociocultural context of learning demonstrates the negative effects of “scarcity” (Mullainathan & Shafir, 2013; Obradovic & Armstrong-Carter, 2021; Verschelden, 2017) on “mental bandwidth.” Goldrick-Rab et al. (2017), for example, report evidence of impact on learners when basic human needs are not met. Two thirds of community college students surveyed in the United States experienced food shortages and 13% or more reported some form of homelessness. Also, over 2% of students in public schools (over 1 million students) in the United States experienced some form of homelessness during the 2020–2021 school year (National Center for Homeless Education, 2022). Such scarcity is likely to impact well-being, attendance at school, and the opportunity to obtain the maximum benefit from available learning opportunities. These adverse social conditions can also impact students’ ability to demonstrate the knowledge, skills, and capabilities targeted by tests.

The implication of this evidence is that accountability decisions based on assessment results should not be based on assumptions that teacher practice is the sole factor involved in student achievement. Narrow interpretations of learning outcomes solely focused on teaching overlook the potential impact of poverty, racism, and social marginalization on cognitive resources (Verschelden, 2017) and the tremendous differences in the sets of opportunities to learn across different groups.

Social Complexity

The way we view the complex relationship between assessment and society is illustrated in Figure 3.2. The diagram has the limitation that it is an overly simple representation of the sociocultural context in the United States and may not entirely reflect the sociocultural contexts of other countries, which may have different but similarly intricate patterns of relationships between assessment and society. Still, it illustrates the reasoning that can be used to examine this relationship.

Three types of interacting components are distinguished: the main stages in the process of large-scale assessment (Components 1–5); the social, legal, and knowledge

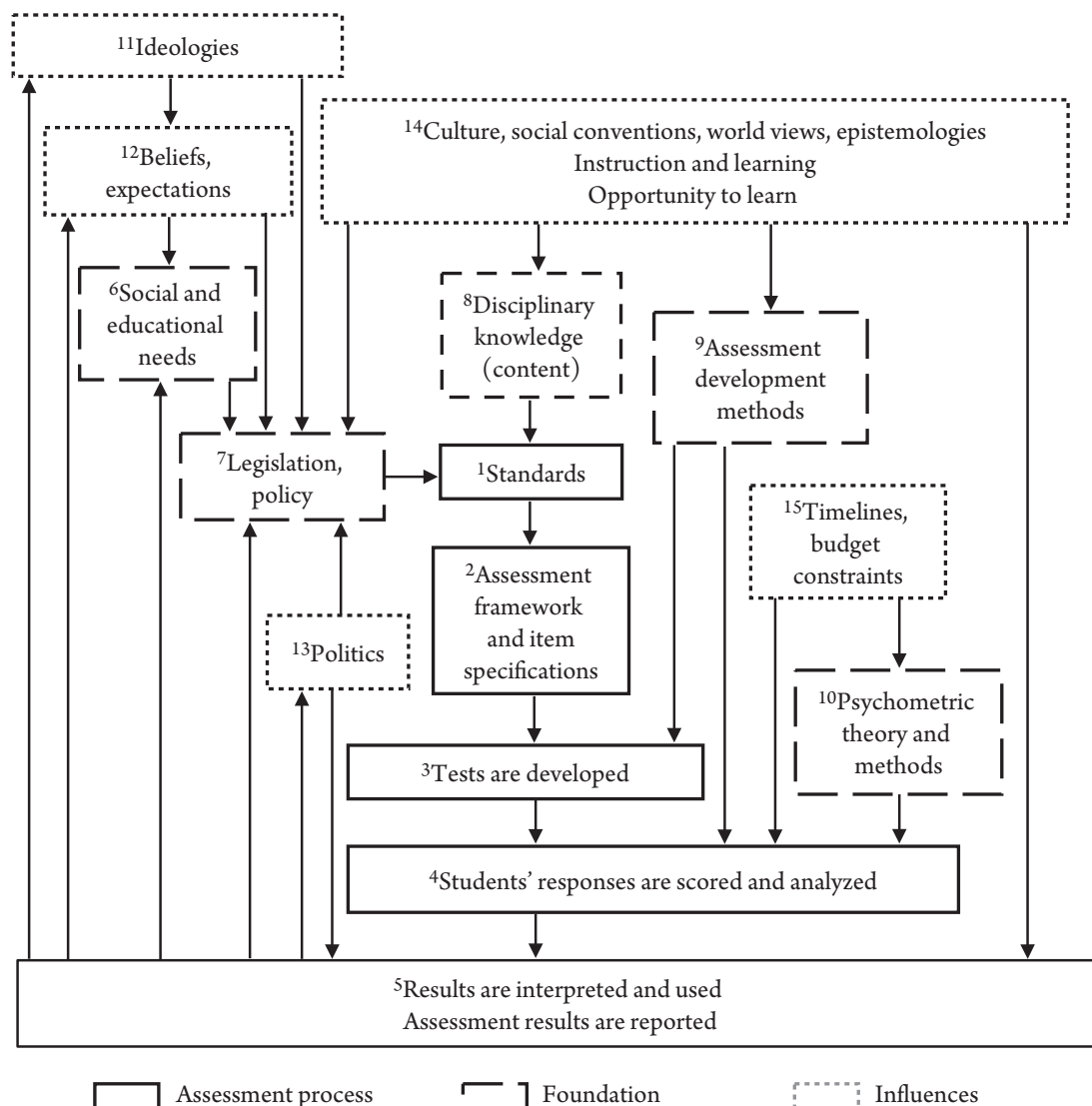


FIGURE 3.2
Relationship Between Large-Scale Assessment and Society

Note. Numbers refer to the components of the relationship between assessment and society, as described in the text.

foundation of assessment as an institutionalized activity (Components 6–10); and the sociopolitical influences and practical constraints that shape the role and impact of assessment in a society (Components 11–15). Figure 3.2 represents the notion that the social and legal foundation of assessment is an institutionalized activity made explicit through formal documents (sometimes legislation) that justify, mandate, and regulate assessment activities. Also, it is supported by a body of knowledge (e.g., current knowledge in a disciplinary field) that determines the content of assessment and by a set of paradigms and methodologies that establish what is acceptable practice concerning assessment development, how assessment data are to be obtained and analyzed, and how assessment results are to be reported.

Social Nature of Assessment

Assessment is assumed to originate from the need to both understand and provide evidence. It is “intended to produce knowledge on the state of an empirical system” (Tal, 2013, p. 1159). All cultures develop measurement strategies that meet societal needs (e.g., Lipka et al., 2015; Owens & Kaleva, 2007). Often, such needs entail construing persons’ differing capabilities in terms of interpersonally well-defined attributes, operationalizing the attributes so construed, and then reasoning through this construal, comparing persons on certain attributes or determining how much of those attributes they have. Validation is the process of marshaling empirical and theoretical evidence to support this reasoning for some persons, contexts, interpretations, and uses. Assessment, as a formal, institutionalized activity, concerns sampling: The tasks or items used in a test can be viewed as samples of observations (Kane, 1982). For a test to be accurate, constructs need to be operationalized into samples of observations that are representative in content and number of the target constructs. Not surprisingly, a significant part of measurement modeling focuses on accounting for measurement error due to the nature and size of the sample of observations (i.e., the content and number of items) used in tests.

Viewing assessment as a social phenomenon (see Vera, 2015) allows examination not only of what is valued in assessment but also of how assessment results are interpreted. For example, test scores can affect an individual’s civic life in several ways. How a person (or their children) has been exposed to schools and tests and the consequences of test scores in life shape their conceptions of educational and testing issues and lead them to develop an opinion on those issues. To the layperson, the causes and implications of a drop in the scores in their children’s school district may be easier to understand (correctly or not) and to act on (effectively or not) than an increase in the average temperature on Earth. The effect of inaccuracy in interpretations of test scores and test score trends are not particular to the general public: Even education officers may advance inconsistent explanations of the decline of national scores in an international test (Morsy et al., 2018)—which may be a reflection of the challenges of examining the role of noninstructional, social factors as sources of measurement error and the perception of tests as instruments that produce precise measures.

Theories of social reproduction (e.g., Bourdieu, 1977; Collins, 2009) provide a lens for examining the complex relationship of assessment and society. As cultural products, tests reflect the thinking, values, and cognitive styles of the persons and institutions that create and use them (Cole, 1999). Institutions and systems in charge of testing infuse their view of the constructs measured and become an important influence on what knowledge is valued and how that knowledge is to be demonstrated. This process can lead assessment systems to prioritize cultures, varieties of language, learning styles, and situations that are aligned with the mainstream culture. This happens even when the number of persons adhering to such practices is no longer a majority, putting individuals who are not from mainstream cultures or in political power (e.g., those who are from lower socioeconomic backgrounds and minoritized economic, racial, ethnic, and linguistic groups) at a disadvantage.

The influence of costs and the technology available for assessment development, administration, and scoring should not be underestimated. For example, large-scale assessments greatly rely on multiple-choice items and make limited use of constructed-response formats because of cost-effectiveness. The former are machine scorable; the latter typically involve human judgment and, therefore, time-consuming activities such as the development of scoring rubrics and the training of scorers (e.g., see Bejar, 2017a; Downing, 2006; General Accounting Office, 1993; Koretz & Hamilton, 2006; Phelps, 2000). While it can be argued that recent advances in artificial intelligence make automated scoring of open-ended responses possible, these possibilities have costs associated with the development of scoring engines and come with the risk of underrepresentation of constructs. Similar limitations are associated with the development of test items using ChatGPT and other generative artificial intelligence technologies and the use of games and simulations as surrogates of hands-on tasks. As a result, cost and development time are important factors in the types of items used and construct representation in large-scale assessment.

Perceptions of Assessment as "Scientific"

In addition to contributing to scientific defensibility, measurement arguments are used to support policies and decisions that affect individuals, institutions, and societies. Student promotion or retention, accountability, school sanctions, and even the ranking of countries are based on comparability and accuracy assumptions (Koretz, 2008). Understanding the sociocultural context in which assessment instruments originate and the operating motivations for their use (and misuse) are central to evaluating these assumptions and moving assessment forward.

How quantitative assessment outcomes influence society by informing decisions about individuals, groups, or institutions is determined not only by the scientific defensibility of these outcomes and the processes that generate them, but also by the views of different segments of the society that utilize quantitative data. According to this reasoning, the effect of tests and assessment on society is shaped not only by the quality of data but also by societal perceptions of what counts as scientific and what makes quantitative data scientific.

Indiscriminately ascribing scientific rigor to quantitative data may contribute to or be associated with different aspects of the complex relationship between assessment and society. One of these aspects may be the lack of a critical interpretation of test scores, which stems from the fact that quantitative information is deemed unconditionally accurate and scientific (Mlodinow, 2008). An example of misuse resulting from this lack of critical perspective concerns international tests.

The view that quantitative data are unquestionably scientific has contributed to the development of polarized perspectives of educational research according to which the qualitative is deemed subjective and the quantitative is deemed objective (Ercikan & Roth, 2006). While psychometrics is concerned with the measurement of psychological, unobservable qualities, efforts to examine the validity of interpretation and use of methods of inquiry on mental process are based on subjects' verbal reports. Perhaps as a reflection of this misalignment, cognitive validation—validation based on the use of think-aloud reports and cognitive interviews to probe how tests elicit the reasoning targeted by tests—was not adopted as a method of validation until the 1990s (Glaser, 1991; Messick, 1995). Still, while there is a wealth of literature on the value of cognitive processes in examining the kind of reasoning elicited by test items (e.g., Ercikan et al., 2010; Ercikan & Pellegrino, 2017; Ericsson & Simon, 1993; Leighton & Gierl, 2011; Ruiz-Primo et al., 2001), methods for cognitive validation using qualitative methods are underused or underreported. It must be acknowledged, however, that thanks to recent developments in information technology, artificial intelligence, and natural language processing, it is possible to obtain quantitative information on test response processes relevant to examining cognitive validity. For example, now it is possible to track eye movements during reading (e.g., Oranje et al., 2017), process verbal responses (e.g., Leighton, 2017), analyze thousands of written essays (e.g., Alvero et al., 2021), and code student responses (e.g., Ercikan & McCaffrey, 2022; Ferrara & Qunbar, 2022) with unprecedented levels of efficiency. Yet while such advances pave the way for research examining the multifaceted nature of constructs, they also put pressure on continually updating of our views of validity (Mislevy, 2018). Rapid adoption of innovative sources of data in assessment without proper support of empirical evidence on validity and fairness may compromise score meaning interpretation (see Bejar, 2017b; Dorsey & Michaels, 2022).

This disappointing scenario is aggravated by the limited scope of the work that examines cognitive activity in assessment. As mentioned before, efforts to examine cognitive validity typically fail to consider the sociocultural context in which cognitive activity takes place. Yet there is evidence that speaks to the importance of examining sociocultural activity as part of the process of test validation. For example, poverty may influence the reasonings used by students when they engage in solving mathematics items that involve addition, multiplication, and rounding (Solano-Flores & Nelson-Barber, 2001). The cultural dimension of thinking and learning is yet to be systematically and routinely incorporated in research and practice in educational measurement.

Unlike the legal and knowledge foundations of assessment, the sociopolitical and practical components that influence the process of assessment are not necessarily explicit and are not formally established. Moreover, being of a dynamic nature subject to negotiation among stakeholders with different objectives and values, these sociopolitical and practical components may be difficult to predict or control. For example, at a given time, beliefs and expectations held by a particular segment of society may lead to inappropriate interpretations of assessment outcomes. Or, under a given political climate, information from tests may be overemphasized or neglected, or the impact of tests on policy may be mediated by political forces.

ASSESSMENT AND SOCIETY: A BIDIRECTIONAL RELATIONSHIP

The purposes and content of assessment, how assessment is conducted, and how assessment results are used and interpreted involve many decisions made by policy makers, test developers, testing agencies, and test users. We argue in the following that each of these choices about assessment both influence and are influenced by sociocultural context.

Societal Needs

The use of assessment with the intent to meet societal needs is far from new. An oft-cited example comes from 14th-century imperial China when, as part of government personnel selection practices, civil and military examinations were used to make hiring and promotion decisions based on rankings on tests (see Brown, 2022; Clauser et al., this volume). However, whether this system of examinations promoted social mobility strictly based on performance is subject to discussion because the emperor had the final word in the selection process of candidates for public service (Elman, 2013). More recent examples of early uses of testing include screening tests for developmentally delayed children in France by Binet in the 20th century (Binet & Simon, 1907) and the adoption of Binet's intelligence test as the Stanford–Binet test in the United States by Louis Terman and colleagues for selecting individuals into the military according to a set of minimal competencies (Gould, 1981). Both instruments involved the classification of individuals. In Binet's tests, in France, the assessment focused on separating children whose development was delayed (as compared to the general population) to provide them with certain services and/or institutionalize them. To measure their abilities, individuals were assessed using basic, everyday life tasks. Views of society and gender influenced the actions taken when the same test was adopted in the United States to meet a very different purpose—selecting men to serve in the military. The concrete life tasks of the test were deemed to be favoring women and therefore “biased,” and decisions were made to include tasks intended to assess abstract reasoning, a construct considered more consistent with expectations for men serving in the military (Gould, 1981). Ironically, while the changes were intended to ensure validity, given the goal

of selecting male army recruits, they may have contributed to promoting narrow and sexist views of intelligence.

A more recent example of how societal needs drive assessment is observed in education reform movements. One motivation for school reform is the need to improve education to create a competitive workforce and reduce disparities in achievement between genders and between ethnic/racial groups. Another motivation is the need to inform decisions concerning the distribution of resources among states, districts, and schools. In the United States, public recognition of discrepancies in the funding of education across districts in the country has led to education reform policies such as the Elementary and Secondary Education Act of 1965 (ESEA) and the No Child Left Behind Act of 2001 (NCLB). The stated goal was to provide equitable financial support and address inequities in educational outcomes. For example, NCLB legislation (NCLB, 2002) involved the development of academic content standards and tests and the establishment of proficiency levels with the intent to monitor and evaluate the effectiveness of education.

Sociopolitical Ideologies

The use of testing for decision-making is often driven by alignment with certain ideologies. The most problematic example of this can be seen in the eugenics movement, which used testing to provide so-called evidence to support racist political ideologies (Gould, 1981). A very different ideology, meritocracy, was grounded in social equality and equal opportunities for individuals from different cultural backgrounds and uses assessment as a key source of information for decision-making.

Among the examples of this use is the inclusion of tests like the SAT or ACT to inform decisions about admission to higher education institutions. While historically there has been a well-supported argument that tests facilitate meritocratic decision-making, more recent discussions highlight the limitations of meritocracy as a singular fairness principle that should inform decisions solely based on tests. Those who support the elimination of the SAT and ACT as part of the set of college admission requirements point to the differential performance of individuals from African American and Latino backgrounds as lowering their chances of admission into higher education, and these supporters of elimination thus hold these tests responsible for the associated inequities. In 2020, the University of California Board of Regents approved a recommendation by Janet Napolitano, then-president of the University of California system, to eliminate the requirement of the SAT and ACT in admission decisions. This recommendation involved making the use of these tests optional for 2021 and 2022 admissions and not using them for 2023 and 2024 admissions for California residents (Anderson, 2020). After the University of California's decision, other universities followed suit (Hubbler, 2023).

Ironically, while arguments in support of eliminating the SAT and ACT are predicated on the premise that these tests are biased against traditionally underserved students, little is said about the fact that the sources of information left to make

admission decisions may be biased too. Biased human judgment may have an especially deleterious impact for traditionally underserved groups in a process of admission that is not informed by any common measure of achievement. For example, applicants who have attended prestigious schools or whose social capital allows them to obtain letters of recommendation from prominent citizens might be at an advantage over other, less privileged individuals. This view is gaining growing support as several institutions started reversing their test-optional, test-free admissions policies (Huebeck, 2024).

The debate on the elimination of standardized tests for college admissions illustrates how testing practices may be shaped by social processes beyond empirically supported arguments. Multiple angles on fairness speak to the complex nature of the relationship between assessment and society and highlight the importance of the current sociopolitical ideologies in shaping the role of assessment in important societal outcomes.

Valued Outcomes

A key decision point in assessment involves choices around what is assessed and how assessments are implemented. What is assessed reflects what is deemed important and relevant for the purposes of testing and the knowledge and skills valued in the society. For example, as we have discussed, assessing abstract reasoning was deemed relevant in the early 20th century in the Stanford–Binet tests for the purpose of selecting men into the military over tasks assessing more practical skills. More recently, emphases on critical thinking, collaboration, problem-solving, and technology and engineering literacy reflect values in the society embracing these constructs as important for success in today’s socio-cultural contexts (e.g., Braun et al., 2020). The surge of interest in performance assessment as a potential tool for assessment-based education reform in the 1990s illustrates these emphases. If performance assessments are more sensitive to higher order skills than conventional, multiple-choice items, then their use in large-scale assessment programs with the purpose of accountability could drive teachers to promote problem-solving strategies and critical thinking among their students (Shavelson et al., 1990).

An example of the influence of sociocultural context and the trade-offs in assessment it motivates was seen in the implementation of NCLB, in what was assessed (primarily reading and mathematics), how students were assessed (primarily with multiple-choice items), and how performance was evaluated (mainly through comparisons against a level of proficiency). The focus on reading and mathematics is a clear example of what was prioritized and what the limitations were of the decisions made: There is a wealth of research on the importance of reading and mathematics for learning more broadly in all subjects. However, excluding science, social studies, and arts from NCLB put teaching and learning in these subjects in second place (Overbaugh & Lu, 2008; Winstead, 2011).

Concepts of Fairness

The concept of fairness has revolved around the notion that performance on tests should not be influenced by factors that are irrelevant to the constructs that tests are intended to measure. This notion has been invoked to evaluate fairness and support decisions

about the content and format of tests, the language in which they are administered, the item types included, and the administration of tests in standardized conditions. For example, the perception of multiple-choice tests as objective because their scoring does not require human judgment and standardization in the ways in which tests are administered stems from the intention of establishing fairness and creating an even playing field for test takers at a time when such conditions were considered sufficient for fairness (Mislevy, 2018). As mentioned, more recent conceptions of fairness include the establishment of conditions that optimize the opportunities made available to test takers to perform on tests, which may be different for different individuals (Mislevy, 2018; Thurlow & Kopriva, 2015). We see clear examples of these conceptions in test accommodations for students with special needs. Since the 1990s, the Americans with Disabilities Act (ADA; 1990, 2008) requires testing organizations to offer tests to persons with disabilities in a manner that allows them to demonstrate their knowledge, skills, and competencies in optimal ways. These accommodations allow item types and formats to be modified with the intent to create equal opportunities for individuals with disabilities to take tests and enhance the fairness of interpretations and uses of assessment results (Abedi & Ewers, 2013). (See also Rodriguez & Thurlow, this volume, for discussion of fairness issues related to students with disabilities and English learners.)

Changes in conceptions of fairness highlight that this fundamental concept in assessment is socially grounded. A relatively recent conception of fairness has been expanded to recognize cultural and linguistic diversity in societies around the world and the implications of such diversity on assessment and measurement (Ercikan & Lyons-Thomas, 2013; Ercikan & Oliveri, 2013; Ercikan et al., 2014; Grover & Ercikan, 2017; Solano-Flores & Li, 2013). Such diversity is officially recognized in some countries, such as Canada, with 2 official languages, and South Africa, with 11 official languages, where most tests are offered in all the official languages.

Official language policies may fall short of fairly considering the complexity of sociocultural context. For example, in South Africa, the number of languages spoken at home is much greater than the number of official languages (Setati & Adler, 2001). Furthermore, these policies do not consider the Indigenous languages. In countries with Indigenous populations such as Canada, Mexico, and the United States, teaching and assessing in Indigenous languages is difficult because of the social stigma attached to them or limitations in the number of educators who are native users of those languages (Oliveri et al., 2015; Simon et al., 2013; Solano-Flores et al., 2015). Recognition of the heterogeneity of cultural groups has led to the development of methodologies for taking such heterogeneity into account when examining fairness (Oliveri et al., 2014).

Researchers have argued that opportunity to learn—“whether or not . . . students have had the opportunity to study a particular topic or learn how to solve a particular type of problem presented by the test” (Husén, 1967, pp. 162–163, cited in Burstein et al., 1993) should be considered part of the efforts to ensure fairness and, more specifically,

the equivalence of score meaning for different groups (Kane, 2012). Opportunity to learn has clear relevance to the appropriateness of interpretations and uses of assessment results in cases in which learning outcomes are used to evaluate teachers and for accountability. In the context of COVID-19, opportunity to learn greatly varied among students from different socioeconomic backgrounds (Kane & Reardon, 2023). Factors that might have limited the opportunity of low-income students to learn remotely during the COVID-19 pandemic started with the level of access to the Internet, including digital devices, familiarity with devices and features for effective learning experiences, schools' and teachers' ability to switch to online learning in a very short period of time, and parents' ability to support online learning at home, among many other possible adverse circumstances. In this context, digital assessments reflect complex interactions between these social factors and the students' knowledge, skills, and competencies.

Taking sociocultural context into account is central to the concept of fairness in the interpretation and use of assessment results. An example of problems that can arise with inferences made from assessment results when opportunity to learn is not taken into account is provided by Shepard (1993). The example is focused on a kindergarten "readiness" test, considered an accurate measure of basic skills, a good predictor of performance in kindergarten, and "fair and valid," using narrow notions of fairness and validity based solely on psychometric criteria. Problems can arise if the test is used to admit children to kindergarten or delay admission because low scores may actually be an indicator of a home environment that does not promote the development of those basic skills (i.e., a lack of opportunity to learn). Holding children back based on test results that are mere indicators of opportunity to learn would be a clear violation of fairness defined as a consistency of score meaning and interpretation of assessment results.

Globalization

Globalization offers many examples of how social contexts provide an important impetus for assessments. The impact of globalization is enhanced through greater sharing and communication of policy and practices in online environments and through greater educational exchanges and expectations for exchangeability of educational outcomes such as professional education programs. The globalization of the labor force has contributed to pushing competitiveness among countries with regard to their educational outcomes, as captured by country rankings based on international tests. Teltemann and Klieme (2016) highlighted the challenges countries face regarding the effectiveness of their education systems as a primary reason for joining efforts in developing international tests of educational outcomes such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).

Technological Advances

A significant societal development that is transforming assessment and learning is the series of advances in information technology and computational tools for using a variety

of data types in assessment (Ercikan et al., 2020; Katz & Gorin, 2016). These technological advances make it possible to use a wide range of assessment types, including interactive tasks, adaptive testing, and open-ended or performance tasks that can be scored automatically using scoring engines. These technological advances are changing not only assessment practices, but also *which* constructs can now be assessed, including collaborative problem-solving and interactive communication. Through the administration of tests in digital environments, we have access to data on student response processes in assessing such complex constructs in ways that are not possible with paper-based tests. Response process data, also referred to as process data, are changing what is considered construct-relevant evidence and how we conceptualize validity and fairness in digitally based assessment environments (see Bennett et al., this volume). Advances in artificial intelligence and other related computer science technologies are utilized increasingly in many aspects of assessment development and scoring of performances. These include the development of test content and item generation and the scoring of text and speech as responses to assessment tasks (see Nikolic et al., 2023; Yaneva & von Davier, 2023).

Unfortunately, advances in information technology, especially the use of artificial intelligence-based algorithms, can be an agent of reproduction of social inequities (Noble, 2018). This concern is important to consider when using artificial intelligence-based algorithms to derive scores either based on responses in formal assessment contexts or using “big data” on individuals’ engagement in online environments such as news, shopping, and entertainment sites (DiCerbo, 2020; King & Mrkonich, 2016). Artificial intelligence-based algorithms use text, images, speech, and test-taker actions to develop features generated from natural language processing in the scoring algorithms. These algorithms are critically dependent on the data sources used in their creation (Baker, 2021; Manyika et al., 2019). Scores may not have equivalent validity and accuracy for all groups if data sources for these algorithms are restricted to specific cultural and language groups (Noble, 2018). For example, on an English-speaking test, the scores from the artificial intelligence-based scoring engine may be biased for students who have different accents or use nonstandard dialects (Benzeghiba et al., 2007). An automated scoring algorithm developed using data from mainstream student groups resulted in less accurate scores for certain racial and ethnic student groups (Bridgeman et al., 2009, 2012). Similar types of bias may be observed in algorithms used for scoring written responses.

An additional challenge in utilizing artificial intelligence-based scoring algorithms using assessment response data or big data is that even though patterns in text or speech or actions may have a specific relationship with the targeted construct, these relationships may vary because of differences in educational and life experiences (Ercikan et al., 2020; Guo & Ercikan, 2021). Inaccuracy in scores and bias may result if the scoring algorithm assumes a uniform relationship between patterns and the targeted construct. Therefore, taking sociocultural context into account has heightened importance when engagement in tests is evaluated using artificial intelligence-based algorithms.

Assessment Capacity

How assessment is conducted is inevitably affected by a wide variety of conditions, including assessment literacy, the availability of individuals with expertise in assessment, and technological resources that can be used to administer assessment and analyze assessment data. Such conditions are undoubtedly expected to affect the quality of assessment and the degree to which assessment can fulfill its intended role.

The context of international test comparisons such as PISA and TIMSS illustrates how certain institutional and human resources are needed for countries to benefit from their participation in these assessments (Husén, 1983). Assessment capacity—the ability of a country to produce assessment instruments with sound psychometric qualities and to develop and sustain its own assessment systems (Solano-Flores & Milbourn, 2016)—greatly influences the extent to which a nation benefits from participating in international comparisons, for example, by being able to make sound interpretations of test scores that inform its educational policy decisions. Poor countries not only tend to perform poorly in international test comparisons, but also have limited resources needed to influence the process of assessment development and take advantage of their participation in international tests as an opportunity to increase their assessment capacity. In addition, this lack of assessment capacity can limit countries' engagement in grounding assessment in their sociocultural contexts. This context of international assessments is a case in which the bidirectional relationship between sociocultural context and assessment is expressed as a vicious circle of inequity at the international level (Solano-Flores, 2019). It also shows, at an international level, a pattern of social inequities and power asymmetry that parallels the social inequities and power asymmetry observed among different racial and socioeconomic groups within a country.

Social Inequities and Test Taking

How assessment is viewed and how human capabilities are valued by society can affect the performance of individuals on tests and how assessment results are interpreted. Two contrasting examples can be used to illustrate the multiple and complex ways in which this relationship between sociocultural context and assessment is expressed. On the one hand, some argue for the motivating effects of the importance given to tests and attribution of fairness and meritocracy to assessment in, for example, East Asian student populations (see Kennedy, 2016). On the other hand, researchers have argued that stereotype threat (fear of confirming a negative stereotype on one's social or ethnic group) may have an adverse effect on the performance of racial minorities on tests (Steele et al., 2002). Structural stigma, the “societal-level conditions, cultural norms, and institutional practices that constrain the opportunities, resources, and wellbeing for *stigmatized* populations” (Hatzenbuehler & Link, 2014, p. 2) is a related concept to stereotype threat; however, it encompasses broader societal contexts that may affect learning and socioemotional and cognitive abilities when learners engage in responding to tests. Even though research has examined the potential of culture as a source of construct-irrelevant variance, mainly bias, this research has had a narrow focus on

cultural references and the linguistic features of items. There is need for more research that investigates the mechanisms through which the broad sociocultural context (which may include the context in which the assessment is conducted as well as individual students' personal contexts) may influence performance on tests and guide appropriate test use and interpretation, given the potential impact of stereotype threat and structural stigma.

Legislation

In the United States, legislation plays a key role as an enforcer of views of assessment and the functions it should have. The effect of the ADA, when it became law in 1990, illustrates the impact of legislation on assessment. The ADA “prohibits discrimination against individuals with disabilities in all areas of public life, including jobs, schools, transportation, and all public and private places that are open to the general public” (see Family Institute, 2024). Since the 1990s, assessment systems have been required to provide accommodations to students with disabilities with the intent of providing them an equal opportunity to demonstrate their knowledge and skills. These accommodations include read-aloud options, Braille versions, and extended testing time options, among others.

Another piece of legislation with impact on assessment has been Title VI of the Civil Rights Act of 1964, which is intended to address the needs and rights of speakers of other languages in public education. This act required districts and schools to provide equal opportunities and services to linguistic minority students in learning and assessment. Optimizing testing conditions for linguistic minority students as policy mandates have also been addressed in this act and have included mandatory provision of bilingual versions of tests, access to dictionaries, and efforts to minimize unnecessary linguistic complexity of the assessment tasks (Takanishi & Le Menestrel, 2017).

Impact of Assessment on Education and Society

Assessment has the potential to impact society in multiple ways. In the United States, assessment has played a central role in providing information about trends in educational outcomes and in shaping teacher practice, student learning, and policies around resource allocations. Assessments have been used as key sources of information for evaluating the effectiveness of education and for holding states and local education agencies accountable, thereby defining expectations for teaching and learning.

In international contexts, assessments of learning outcomes such as TIMSS or PISA receive both positive and negative attention in participating countries, which often leads to significant changes in policy and practice. When participating countries were surveyed about the impact of PISA on their education systems, more than half of the countries reported PISA as being “extremely influential” or “very influential” (Breakspear, 2012). Discourse and rationales for reform efforts and policy changes have used PISA or TIMSS in justifying the direction for policy changes (Feniger & Lefstein, 2014). In Canada, rationales for reform have included the alignment of within-country assessment programs with PISA and the harmonization of national indicators with the

Organisation for Economic Co-operation and Development's at-a-glance indicators (Froese-Germain, 2010).

Another impact of assessment has been the shaping of education policy and practice. We see examples of this impact in England and Germany, which have made major changes to teaching practices, teacher training, instructional materials, educational standards, and curricula based on information and insights gained through international assessments (Heyneman & Lee, 2014). The comparative emphasis in international assessments shifted the focus from ideology-based policies to evidence-based policymaking based on assessment results (Bieber et al., 2014; Grek, 2012), as well as a shift from "input oriented" to "output oriented." At the root of this comparative perspective is an assumption that education systems are comparable and policies and practices are transferable. Despite criticisms of this assumption by researchers (see Carnoy & Rothstein, 2013; Ercikan et al., 2015; Singer et al., 2018), a consequence of that perceived transference is a promotion of *convergence of education systems* toward each other and globalization of education (Drezner, 2001).

When developing an understanding of the impact of assessment on society, examination of the unintended consequences of assessment is as important as examination of the intended consequences. Intended consequences of assessment are easier to understand and are likely to be more openly invoked by decision-makers and policy makers. Obvious examples include changing education practice to improve learning, driving policy to narrow disparities in educational achievement, and making changes in professional education that may contribute to bettering health outcomes or reducing societal risks of malpractice in professions.

Unintended consequences of assessment may be more difficult to recognize, but are important to consider in improving education systems. A clear example is the impact of statewide assessments used for accountability purposes on Indigenous populations in the United States. There is evidence that the pressure to meet demands imposed by NCLB drove some teachers to neglect instruction based on firsthand, meaningful cultural experience and to instead emphasize instruction oriented to promote high performance on statewide summative assessments (Trumbull & Nelson-Barber, 2019).

Assessment of individual competencies inevitably reflects educational opportunities, practices, and resources, which are all affected by inequities in societies. When assessment results inform high-stakes decisions (e.g., the admission of individuals into higher education institutions), these uses may contribute to perpetuating educational and social inequities. Other assessment uses with potential for unintended consequences include their inclusion in high-stakes accountability decisions for rewarding or penalizing schools and teachers and for allocating resources for schools. In such scenarios, schools and teachers working in the most challenging educational contexts may be unfairly evaluated by formal accountability models that are based on value-added models or simple comparisons of performances across schools.

LEGISLATION AND ASSESSMENT IN THE UNITED STATES: A STRONG BIDIRECTIONAL RELATIONSHIP

Legislation and policy have shaped assessment and its role in education and the society in significant ways, which speaks to the strong bidirectional relation between sociocultural context and assessment. Laws have been mechanisms through which many aspects of the sociocultural context have found ways to shape education and assessment, including expectations for teaching, learning, and allocation of resources in education. In fact, assessment has been used as a key tool for implementing significant legislation since the 1980s.

Legislation has played a central role in shaping assessment in two related ways. One is when legislation dictates features, uses, and interpretation of assessments; the other is when it is intended to prevent bias and discrimination to protect individual rights.

In U.S. educational contexts, the foundations of the rules for how assessments are conducted and used are grounded in the U.S. Constitution, federal civil rights statutes, judicial decisions, and state laws. Current federal, state, and local legislation and policies use student assessment for eight distinct purposes, with the same test often serving multiple functions (NRC, 1999):

- evaluating and improving district-level educational programs;
- providing information about individual students to inform instruction;
- monitoring educational outcomes and providing information about the status of the education system;
- motivating change in education;
- informing program evaluation by using pre- and posttest results to gauge the impact of educational programs;
- holding states, local education agencies, schools, and educators accountable for student performance;
- promoting change in classroom instruction; and
- certifying student achievement as meeting a performance standard, typically used in graduation from high school or promotion to the next grade.

The legal context of assessment is particularly relevant to high-stakes testing. Practices and instruments may be challenged when there are claims of misuse or discrimination. In state tests, there are four major areas of concern: (a) testing to award diplomas; (b) potential bias against historically disadvantaged groups; (c) providing testing accommodations to individuals with disabilities; and (d) issues specifically related to performance tests.

In this section, we describe how these legal contexts shape assessment in the United States, as well as how assessment is used to make changes in education and society through legislation. We first describe key legislation that impacts assessment in the United States. We then describe how legislation is used to ensure the inclusion of two populations of students—students with disabilities and ELs—in education and assessment.

Key Legislation Impacting Assessment in the United States

Key federal pieces of legislation relevant to testing in the United States are the Family Educational Rights and Privacy Act of 1974 (FERPA), Section 504 of the Rehabilitation Act of 1973, the ADA Act of 1990 as amended in 2008, the Individuals With Disabilities Education Improvement Act of 1997 (IDEA) and its 2004 reauthorization, and the ESEA of 1965. These five laws inform key concepts in testing and address issues ranging from confidentiality and privacy to discrimination and access to educational diagnosis and overall educational accountability (Borreca et al., 2013).

FERPA is a U.S. federal law that has been in existence since 1974, and several amendments have been made since that time. It involves three issues: access to education records, opportunity to have records amended, and disclosure of information from records. FERPA regulations have been used to access and review tests or amend scores from tests when the results have been successfully challenged (Borreca et al., 2013).

Section 504 of the Rehabilitation Act of 1973 and the ADA of 1990 are intended to protect the rights of children and adults with disabilities. Section 504 is considered the first civil rights statute for persons with disabilities. It is designed to eliminate discrimination based on disability in programs and activities receiving federal funding and covers K–12 public education and higher education, including educational assessment. The ADA was signed into law in 1990 and extends law into private-sector programs and activities that do not receive federal funding. Both Section 504 and ADA relate to educational testing in three areas: evaluation procedures, determination of disability, and provision of accommodations. Evaluation refers to specific processes and tools used for evaluating targeted knowledge and skills of individuals with disabilities, so that their performance accurately reflects such knowledge and skills, by removing the confounding effect of the disabilities. For individuals with disabilities to be protected by Section 504 and the ADA, the individuals need to demonstrate that they have a physical or mental disability that substantially limits one or more major life activities.

In 1965, President Lyndon B. Johnson signed the ESEA. This legislation signaled a significant federal initiative for improving “quality and equality” in K–12 public education. The law was intended to provide substantial resources to support educational opportunities for students from lower socioeconomic backgrounds. The support for public education by the federal government was accompanied by an emphasis on accountability for schools. Initial accountability efforts focused on how schools used the resources for improving education. In the subsequent reauthorizations of ESEA, accountability evolved into outputs, primarily focusing on student learning outcomes. The 1994 reauthorization, called the Improving America’s Schools Act, highlighted a focus on accountability. The biggest shift to accountability through assessment was in the 2001 reauthorization of the ESEA legislation, the NCLB (2002). The associated regulations of NCLB were passed in 2003, and new regulations were added, some of which were rescinded. In addition to research-based instruction and parental choice, accountability was central to NCLB, which required the development of content-area standards for

reading, mathematics, and science, with annual testing of student proficiency in these areas in Grades 3 through 8 and once between Grades 10 and 12. Student performance results were used to evaluate the effectiveness of schools and teachers. NCLB mandated that results be broken down by racial and ethnic groups, for students with disabilities, for students who were poor, and for students with limited English proficiency. Two types of changes were allowed for students with disabilities and those with limited English language proficiency: accommodations (adjustments made to assessments that are not expected to alter the assessed construct) and modifications (changes in test content that are expected to affect the construct being measured). Reporting of adequate yearly progress, an estimation of change in percentage of students performing at *Proficient* or higher levels, was central to evaluating school performance.

The influence of the IDEA of 1997 and its 2004 reauthorization on the field of assessment in the United States cannot be overstated. As a result of this law, there have been significant changes in how assessments are used, developed, and administered. There may be both a positive side and a negative side to these changes. First, requiring students to be evaluated to inform special education referral decisions has contributed to giving assessment a prominent place in educational practice. Second, requiring the inclusion of students receiving special education services in state and district assessments for states to receive their funding for school-age special education has strengthened the role of assessment as a tool for state control. At the same time, it has contributed to promoting a socially shared view of inclusion as both a reachable goal and an institutional responsibility. Third, this requirement of inclusion has attracted the attention of researchers and practitioners to testing accommodations as a viable set of resources to be used by students when demonstrating their knowledge.

At the same time, it must be acknowledged that legislation prescribing inclusion in assessment is not always aligned with the actual advances in the field of educational measurement. As is the case for legislation relevant to ELs (see Solano-Flores & Hakuta, 2017), legislation intended to serve ELs appears to be driven by high expectations about the precision of the information provided by assessments and unrealistic expectations about the ways in which the information is used. Also, while school districts may be diligent in taking the actions prescribed by the law to include students with disabilities through the use of testing accommodations, there may not be empirical evidence of the effectiveness of these accommodations, or they may not be implemented with fidelity. In the end, while the ultimate intent of IDEA is to provide appropriate forms of support to students with disabilities, overreliance on assessment can sometimes drive efforts toward this goal in the opposite direction.

The Every Student Succeeds Act (ESSA; the current authorization of the ESEA) was signed in 2015 and continues to focus on assessment and accountability. Compared to NCLB, ESSA provides more flexibility in how assessment is used in accountability models. However, ESSA continues to require testing of all students in English language arts and mathematics in Grades 3–8 and once in high school. In addition, states are required to test students at least three times in science, once in Grades 3 through 5,

once in Grades 6 through 9, and once in Grades 10 through 12. The flexibility of ESSA includes the possibility of replacing the high school tests with “nationally recognized” college-readiness tests and the use of multiple interim tests in place of a single summative assessment.

Since the early 1970s, these legislative acts have had a great influence on many aspects of education, assessment, and testing. Borreca et al. (2013, p. 450) summarized the implications of these laws and regulations for testing as follows:

- Professionals who work in education at all levels must remain knowledgeable about laws and keep up to date with revisions made through amendments and reauthorizations.
- Educators must ensure that all evaluations are conducted in a manner consistent with federal laws.
- For individuals identified with a disability, educational tests must provide accurate measurement of the student’s knowledge and skills, rather than be adversely impacted by the student’s disability.
- The confidentiality of all testing records must be safeguarded, and test preparation protocols should not be destroyed.
- Tests must undergo constant scrutiny to ensure that they have adequate content-related validity evidence.

The wide range of influences of legislation included expectations for professionals, how assessments are conducted, how individuals with disabilities are assessed, how assessment information and data are stored, how assessment results are to be used, and what psychometric requirements should be met.

Legislating Inclusion of Students With Special Needs

Definition of Students With Special Needs

For the purposes of our discussion, we use the term *students with special needs* to refer to ELs (students who are not proficient in the language of testing) or to students with disabilities. While the two groups of students are very different and the legal provisions for their inclusions in tests vary, there are important commonalities that stem from the challenges posed by the groups’ heterogeneity. In the United States, legislation that mandates the inclusion in large-scale testing programs of students with special needs is intended to ensure that tests produce information on student achievement that ultimately informs policy and practice in ways that benefit all students. This legislation focuses on (a) identifying, through diagnostic testing, those individuals with given characteristics who require certain forms of support intended to ensure that they benefit from instruction and (b) ensuring that individuals already identified as having that condition are tested with supports intended to minimize the effect of that condition as a source of measurement error. This legislation generates a tension between what assessment programs, school districts, and educators are legally required to do and what is technically possible to accomplish.

An important aspect of legislation concerning inclusion is the definition of students with special needs. In accordance with Sections 300.304 through 300.311, “child with a disability” means a child evaluated as having an intellectual disability, a hearing impairment (including deafness), a speech or language impairment, a visual impairment (including blindness), a serious emotional disturbance (referred to in this part as “emotional disturbance”), an orthopedic impairment, autism, traumatic brain injury, any other health impairment, a specific learning disability, deaf-blindness, or multiple disabilities and who, by reason thereof, needs special education and related services. In the case of ELs in the legislation, the ESSA (2015) defines them as “students whose English language proficiency affects their ability to meet state academic standards.”

Testing Accommodations

Legislation for inclusion in assessment largely depends on the use of testing accommodations—adjustments to the ways in which students are tested with the intent to minimize construct-irrelevant variance due to disability or limited proficiency in the language in which tests are administered (see Abedi, 2012; Sireci et al., 2005). Requirements for accessibility are intended for tests to accurately reflect targeted attributes and abilities. In the case of students with disabilities, the examination

is selected and administered so as to best ensure that when the examination is administered to an individual with a disability that impairs sensory, manual, or speaking skills, the examination results accurately reflect the individual’s aptitude or achievement level or whatever other factor the examination purports to measure, rather than reflecting the individual’s impaired sensory manual or speaking skills, except where those skills are the factors that the examination purports to measure. (Department of Justice, 2010, § 36.309)

In the case of ELs, ESEA legislation holds schools accountable for demonstrating improvements in students’ English language proficiency each school year, as well as demonstrating improvements in achievement levels as expected of the other students. Consequently, ELs are required to take English language proficiency tests in addition to taking the same state tests (with accommodations) as the other students. Unfortunately, while these requirements are based on the premise that test accommodations are effective, most test accommodations for ELs are limited in their effectiveness for removing English proficiency as source of construct-irrelevant variance, and the fidelity with which they are implemented is rarely documented (Solano-Flores, 2022; Wolf et al., 2008). Ironically, while accommodations are intended to ensure fairness in testing, the assumption of effectiveness implicit in legislation that mandates the inclusion of ELs in large-scale assessment produces a complex set of unfair conditions for these students.

Limits of Inclusion

The inconsistency between the ways in which special populations are defined and the ways in which research, practice, and ideology construe these populations is a major

factor shaping the complex relationship between sociocultural context and assessment. For example, actions taken by practitioners may be oriented to comply with the law. However, the skills targeted by legislation may not exactly match the constructs identified by research as those that are measured with defensible accuracy. Consequently, efforts to use assessments as policy tools for making changes in schools and improving education may be limited in their effectiveness in improving education.

As with any other legislation, legislation of fairness leaves room for interpretation, which in some cases may lead to undesired outcomes. Educators and test users need to be aware of multiple issues concerning legislation for fairness because decisions about students may have long-lasting effects on students' lives. The first issue concerns the fact, previously mentioned, that legislation that establishes that students must be given diagnostic tests implicitly assumes that tests are more precise instruments than they really are (see Solano-Flores & Hakuta, 2017). Legislation is silent about the false-positive and false-negative identification of students in diagnostic tests. Because measurement error can be minimized but not eliminated, any test may erroneously identify an individual as having or not having a condition (Durán, 2008). For example, in the case of proficiency in English as a second language, one factor that contributes to this challenge is that language is a vast domain. Proficiency in a language is largely dependent on the contexts in which language is used, and these contexts may not be appropriately considered in the design of tests used to make inferences regarding language skills. Also, many tests used to measure English proficiency focus on discrete language skills (Valdés & Figueroa, 1994) and mostly on vocabulary (García et al., 2006) which are far from being sufficiently representative of the multiple aspects of language that are relevant to learning in school (Wolf et al., 2008). Historically, the development of English as a second language has been erroneously viewed as the learning of a subject (Valdés, 2004, 2018) and proficiency in English as a second language at large is often confused with proficiency in English language arts (which involves formal instruction in using language for specific academic purposes). Fortunately, the development of English development standards that are sensitive to learning in school contexts (see Gottlieb et al., 2007) has the potential to contribute to changing that trend.

The second issue has to do with the categories used for classifying students with special needs. In the field of special education, authors have expressed concern that while the disability categories established by IDEA may serve the purpose of identifying who qualifies to receive special education services, the categories are overly broad and overlapping. Concerns also exist that these categories do not carry the meaning relevant to informing decisions on the optimal ways of supporting students with disabilities and may be interpreted differently by different individuals (MacMillan & Hendrick, 1993; NRC, 1997). This mismatch poses serious challenges to both effective identification and valid, fair interpretation of assessment results.

The third issue is the fact that legislation on fairness appears to be based on unrealistic assumptions about the challenges of serving students with special needs and the capac-

ity of school systems to properly implement this legislation. For example, the ESSA authorizes states to test EL students in their first language, guided by the premise that all EL students are more able to demonstrate their knowledge in their first language than in English (Solano-Flores & Hakuta, 2017). In actuality, most EL students have been schooled in English-only programs and have been formally taught to read and write only in English (American Federation of Teachers, 2002). For the majority of EL students, testing them in their first language could be more harmful than beneficial. ESSA is also silent about the nuances of assessing different content areas in the students' first language. For example, is it possible to assess English language arts skills in a different language? As discussed previously, although they are related, English proficiency and proficiency in English language arts are very different constructs. Furthermore, this legislation appears to assume that all states and school districts have the same capacity to translate tests in ways that ensure the comparability of scores across test language versions. Yet a wealth of evidence shows that, in the absence of a strong process of test translation, translated tests may assess different constructs, or the scores produced by tests in different languages are not comparable (Gierl & Khaliq, 2001; Hambleton, 2001; Sireci, 1997; Solano-Flores, 2012; van de Vijver & Poortinga, 1997).

Altogether, lack of consideration of these issues may have a deleterious impact on the sets of opportunities made available for students to progress academically. They speak to the need for extensive training and professional development that enables education administrators and educators to make sound interpretations of legislation and its impact on instructional practices. For example, there is evidence that ELs who are native users of Spanish are less likely than ELs who are native speakers of other languages to be reclassified as English proficient (Faulkner-Bond, 2016), which suggests that there is a bias in the ways in which information from tests, in combination with other sources, is interpreted to make decisions concerning the reclassification of students as English proficient. Also, there is evidence that students from traditionally underserved groups are overrepresented in special education programs (Artiles et al., 2010), which is an indication that there is a bias against these students in the diagnostic assessment of disability. Particularly for EL students (see Abedi, 2014), this pattern also indicates that, despite the information on English proficiency from mandated tests, special education referral decisions may be influenced by erroneous perceptions of students' proficiency in English that educators may develop based on natural aspects of language (e.g., unfamiliar patterns of spelling mistakes and pronunciation) that may be confused with intellectual disability (Klingner & Harry, 2006). Clearly, legislation concerned with fairness cannot have the desired impact in the absence of serious, coordinated efforts that are oriented to promoting the proper interpretation of legislation and improved assessment-use practices.

Unfortunately, given the limitations of existing legislation on assessment and fairness and in a climate of accountability, assessment developers, school officials, administrators, and educators may reach a point where their efforts are focused more on complying with legislation on the assessment of students with special needs than on addressing each special education student's needs. Indeed, an obstacle that may hamper the ability

of assessment systems to properly provide accommodations for students with disabilities is that disability legislation (IDEA, 2004) imposes a system of disability categories that may not serve the unique set of needs of each special education student with the same effectiveness with which it serves the need for operationalizing students for legal and accountability purposes.

In sum, reflecting the concerns discussed, legislation concerning fairness in assessment does not necessarily guarantee implementation of fair assessment. Proper implementation, deep integration of the sociocultural context in assessment design and use, institutional commitment to fairness, and a critical interpretation of legislation are all critical for decision-makers and educators to meet the goals of that legislation.

Legislating Change in Schools and Classrooms: Accountability

Assessment has been used by elected officials and other policy makers as one of the most powerful tools for influencing and changing what happens in schools and classrooms. There is a great degree of evidence that this use of assessment has, in fact, met the intended purposes (Corbet & Wilson, 1991; Durán, 2008; Gándara & Baca, 2008; Herman & Golan, 1993; Milner et al., 2012). These influences include the prioritization of learning outcomes and instructional methods.

The effects of such tests have been debated and have not always been in the direction of improving learning (Mehrens, 1998; Milner et al., 2012). However, compared to other interventions, such as improving the quality of teaching or providing better quality resources, tests are considered relatively inexpensive and quicker ways of influencing what goes on in classrooms and have become the hortatory go-to strategy for policy makers.

Assessment experts and others concerned about the effects of inappropriate uses of tests have cautioned against their use for high-stakes policy goals. Several areas of concern have been highlighted, including creating greater differences in educational opportunities and resources for students from different sociocultural backgrounds, narrowing curriculum, removing local control of educational decision-making, and reducing teacher control and professional decision-making (Amrein & Berliner, 2002; Haertel, 1989; Huebert & Hauser, 1998; Jonsson & Leden, 2019). (See also Ho & Polikoff, this volume, for discussion of assessments used for accountability purposes.)

The use of assessment for such policy strategies has two key challenges. One is related to unrealistic expectations for assessment to meet the intended policy goals. The second challenge is related to the tension between the use of assessment to provide “objective” and “fair” information about learning outcomes and the limitations in creating tests for fair and valid interpretation of assessment results for diverse student populations (NRC, 1999). The unrealistic expectations for tests include their use for multiple purposes, some of which they were not designed for, and thus do not support all the purported interpretations and uses. A common challenge emerges, as a result of financial constraints and timing-related issues, when existing tests are used instead of developing tests for valued, targeted purposes. For example, problems arise when tests that are intended to provide group-level results, such as at school or classroom levels, are used

to make decisions about individual students, such as for grade retention or assigning individual school grades. These problems become expanded and compounded further when what is assessed does not correspond to the opportunities made available for students to learn in their sociocultural contexts.

There is extensive reliance on tests and test results as objective sources of information about student populations, especially highly diverse student populations. Tests have been used for classifying students into achievement levels, for assigning them to classrooms, and for tracking purposes. There are extensive current and historical examples of such uses. During the growth of mass public education between 1870 and 1890, student enrollments doubled in size because of an increase in immigration, which created a more diverse student population. Managing schools during this time required “the objective and efficient classification, or grading of pupils” (Tyack, 1974, p. 4). Tests were believed to have provided fair and objective information for making decisions about these students (Haney, 1984; Office of Technology Assessment, 1992). Researchers have highlighted the limitations of the notion of objectivity and the generalizability of claims based on assessment results for diverse student populations (Bailey & Durán, 2019; Ercikan & Roth, 2009; Ercikan & Solano-Flores, 2014; Solano-Flores, 2009). For assessments to support valid interpretations of scores for diverse populations, this diversity needs to be considered at all stages of assessment design, development, interpretation, and use (Solano-Flores & Nelson-Barber, 2001). However, “taking diversity into account” has many possible interpretations. Research and new conceptual developments are needed to establish how this is to be done in optimal, cost-effective ways.

Foundational Impact of Court Cases on Assessment

Court case rulings have historically had a foundational impact on assessment in practice. There are different types of arguments that are used in court cases for arguing for a particular use as being legally discriminatory in assessment. These include (a) presenting the intent to discriminate, (b) preserving the effect of prior illegal discrimination, and (c) causing a disproportionate adverse impact (Borreca et al., 2013; Phillips & Camara, 2006). These arguments focus on whether tests are used in a discriminatory manner for tracking, promotion, graduation, or licensure and certification.

Intentional discrimination or the intention to discriminate is forbidden under the equal protection clause of the Fourteenth Amendment. This clause forbids schools, districts, or officials from engaging in acts of intentional discrimination based on race, color, national origin, or sex (Personnel Administrator of Mass. v. Feeney, 1979; *United States v. Fordice*, 1992). Lawsuits arguing intentional discrimination have been rare and are difficult to support. These lawsuits have the burden of demonstrating not only the disproportionate adverse impact but also the intentionality of such impact. Regarding high-stakes testing, courts have almost always dismissed claims of intentional discrimination (NRC, 1999).

Lawsuits arguing that the test preserves the effects of prior discrimination are supported by the Supreme Court ruling that the Constitution forbids practices that serve to preserve or continue the effects of prior illegal school segregation (Borreca et al.,

2013). This legal claim was more common in the 1970s and 1980s because attendance at illegally segregated schools is rare to nonexistent in the early 21st century.

In arguing for disparate impact, plaintiffs need to demonstrate that some policy or practice, such as the use of a test, has disproportionate adverse impact on a particular group. This claim cannot be supported by just the difference in test scores for groups. There are two key requirements for disparate impact claims. First, the difference in performance between groups is *too large*, although it is possible to have psychometrically sound instruments that produce very large differences between groups. Criteria for deciding which differences are considered too large involve *practical significance* and *statistical significance*. Practical significance requires the demonstration of a disproportionate number of individuals from a particular group being assigned to treatments or given access to placement opportunities. The second requirement is the demonstration of lack of psychometric defensibility of the instrument or practice, as is the case with insufficient validity evidence for the proper interpretation and use of assessment results. The decisions about such claims may also depend on whether test scores are the only source of criteria in making placement or treatment decisions.

Rulings in the court cases involving due process challenges to some of the high-stakes tests constitute examples of foundational impact of assessment. The landmark *Debra P. v. Turlington* (1979) court case involved a challenge by a group of African American students who were denied a high school diploma based on a score on a statewide Florida competency test. The challenge claimed that the testing requirement was racially biased. The 1979 claim was supported by the discrepant failing rates: 2% for White students versus 20% for Black students. In this court case, the U.S. Court of Appeals made opportunity-to-learn central to fairness. It specifically ruled that if a high school graduation test covers material not taught to the students, then it is unfair and violates the Fourteenth Amendment of the U.S. Constitution. The challenge resulted in two major requirements for fairness: *adequate notice* and *curricular validity*. Adequate notice requires clear communication of the expectations for the test several years before the test is administered to allow students and their associated schools to prepare for the test. Curricular validity requires documented evidence that the students had reasonable opportunities to learn the content and skills covered by the test. The ruling highlighted the importance of connections between schools' accountability in providing opportunities to learn and denial of a high school diploma to students. Since demonstrating actual opportunity to learn at the individual student level is difficult, the U.S. Court of Appeals accepted evidence of the content being included in the curriculum and endorsement of the content by teachers as being skills they would teach (*Debra P. v. Turlington*, 1983, p. 186).

This challenge established new legal requirements for defensible high school graduation tests and other high-stakes assessment uses. First, in designing new tests for high-stakes decision purposes, the risk of not being able to meet the curricular validity requirement arises if sufficient time and resources have not been provided to schools for making the appropriate changes to curriculum and instruction. This is an important consideration when new tests are introduced for high-stakes decision-making purposes. New high-stakes

tests need to allow an adequate amount of time and dedicated resources for the required changes to align curriculum and instruction to the targeted assessment. This may require several years, depending on the level and types of changes that need to be made.

A shortcut to meeting the curricular validity requirement might be thought to be teaching to the test. However, ironically, this strategy compromises curricular validity because of the impact this practice has on learning more broadly and because it compromises the interpretation of test results as an indicator of student learning (Linn, 1983; Popham, 2001). Therefore, simply teaching to the test cannot be used to support curricular validity. There is a requirement to demonstrate that opportunity to learn, and therefore curricular validity, is established without compromising the validity of the measurement of learning.

Standards for testing, such as the *Standards for Educational and Psychological Testing* (AERA et al., 2014), play an important role in the defensibility of tests and test results and are often invoked in legal cases involving testing. They are often cited by experts, plaintiffs, and testing organizations for appropriate measurement, interpretation and use, and consequences of use issues. Central to the defensibility of tests is following professional standards in all technical matters, including, but not limited to, item development, item selection, validity, reliability, item bias review, equating, scaling, setting performance standards, test security, accommodations, test administration, scoring, and score reporting.

DESIGNING ASSESSMENTS GROUNDED IN SOCIOCULTURAL CONTEXT

This section discusses how considering the sociocultural context of assessment contributes to informing and guiding educational measurement decisions. In the following subsections, we first present a discussion of measurement as reasoning from evidence, followed by an exploration of what it would mean to situate this reasoning in the sociocultural context.

Assessment as Reasoning From Evidence Grounded in the Sociocultural Context

Our discussion in this section builds on the three interrelated core aspects of reasoning from evidence referred to as the *assessment triangle*, originally proposed by the NRC's flagship book *Knowing What Students Know* (NRC, 2001). First, we discuss the assessment triangle and then we discuss each of its components from a sociocultural perspective.

Underlying an assessment are three (implicit or explicit, formal or informal) interacting models: a *cognition model* (characteristics of the target construct), an *observation model* (features of tasks and observations expected to provide evidence of the target construct), and an *interpretation model* (rules, guidelines, and statistical models for interpreting observation as evidence of the target construct). These three models need to interact coherently to support assessment claims. Thus, their relationships can be represented as a triangle in which each vertex is intended to inform its two adjacent vertices (see Figure 3.3).

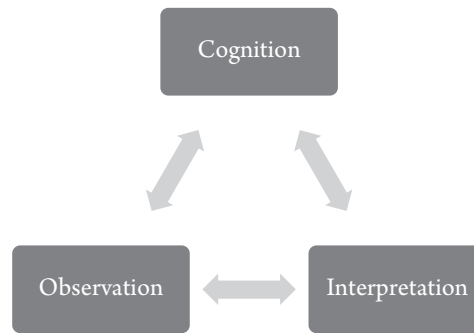


FIGURE 3.3
The Assessment Triangle

We point out the issues and dilemmas that need to be resolved if assessment is to effectively consider sociocultural context. We make a case for each aspect of the assessment triangle—cognition, observation, and interpretation—as well as the relationships among these vertices of the triangle to be grounded in the sociocultural context. Also, we highlight the importance of conditional inference discussed in the first section of the chapter in interpreting and making claims about assessment results (see Figure 3.4).

Cognition Model

The cognition model, also referred to as the *cognition and learning model* or the *student model*, concerns the construct targeted by an assessment and specifies an expected progression in learning. It includes aspects of the construct including knowledge, skills, and competencies in different stages of development and progression in a particular area of assessment.

Two challenges in developing a model need to be acknowledged as influencing the effectiveness with which it is specified. One is representational: A cognitive model postulates a structure of knowledge, for example, in terms of the sequence with which concepts are acquired and their hierarchical or reticular organization (see Nichols et al., 2017). This structure needs to be efficiently represented for proper analysis. For example, it should be able to allow examination of multiple sequences of learning—which can be many, even for seemingly small knowledge domains (e.g., Falmagne et al., 2006).

The second challenge is the difficulty of handling ontoepistemic heterogeneity—the condition that “knowing and being are inextricably tied” (Warren et al., 2020, p. 278). Individuals with different backgrounds (histories, identities, social practices) construct knowledge and make sense of experience differently (Rogoff, 1998). The differences in sociocultural contexts of learning point to possible differences in learning and cognition models, as well as in the organization of knowledge structures for individuals from different contexts (Rogoff, 1998).

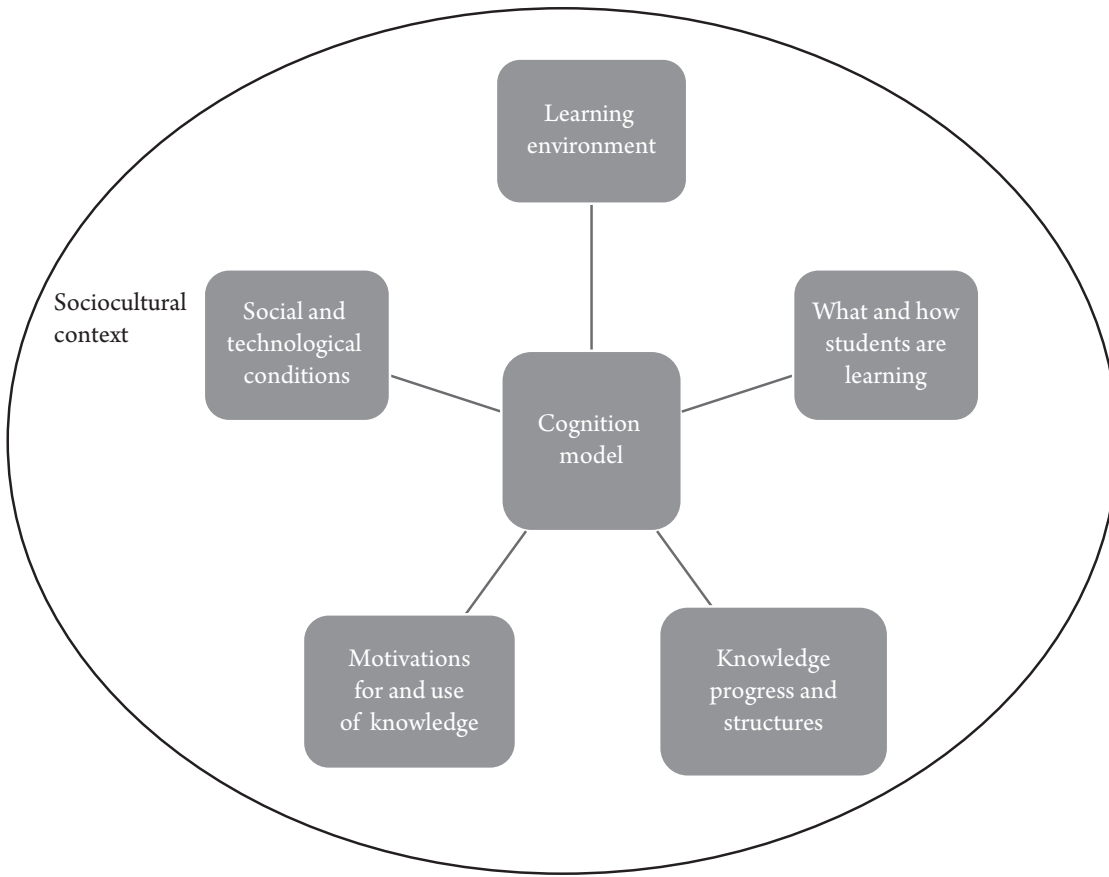


FIGURE 3.4
Factors and Considerations Impacting the Cognition Model

Taking the sociocultural context into account in the cognition model requires gaining insights about what, why, and how students are learning (Donovan & Bransford, 2005; Engeström, 2001), the relationship between the learner and the context of learning, and the conditions of learning. The following are possible guiding questions for examining the equivalence of cognition models for students from different sociocultural backgrounds:

- What are the key aspects of the students' learning environments, including home and schooling environments, technology access, and integration?
- How is the intended learning expected to progress?
- How do the target learners organize aspects of the knowledge they are attaining?
- What are the factors that drive their motivation, or the lack of it?
- How are students expected to apply their knowledge in their everyday lives and at school?

These questions (represented in Figure 3.4) need to be explored through empirical research focusing on differences between students from different sociocultural back-

grounds with respect to how they think and learn, how performance at different levels of proficiency varies, and how they understand and relate to the subject matter.

The equivalence of cognition models is central to equivalence of measurement that allows comparisons of scores across individuals from different sociocultural groups (Ercikan & Por, 2020). This equivalence refers to the degree to which individuals are expected to develop and progress on the target constructs in similar ways and the extent to which the constructs are accessible to students from different backgrounds in similar ways. Differences in some components of the cognition model may require an assessment to focus on parts of the construct that, according to empirical evidence, have been shown to be equivalent across the considered groups.

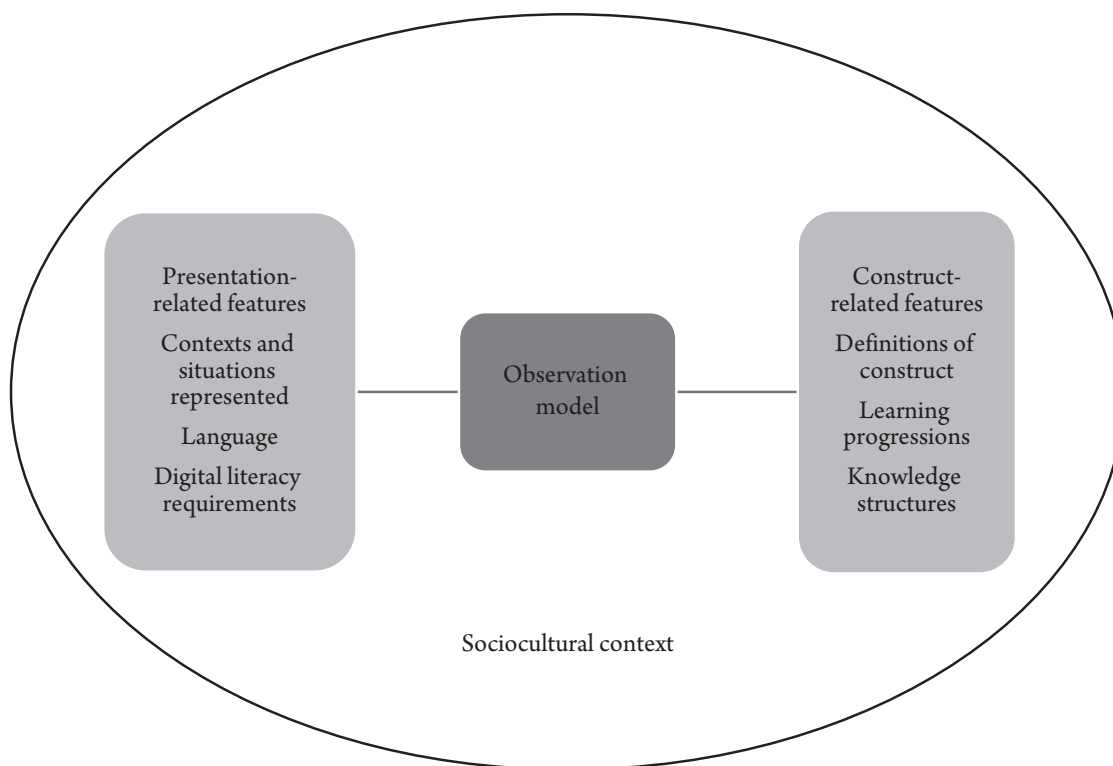
Notice that the emphasis on the ways in which knowledge is acquired may vary depending on the purposes of testing because individuals may acquire knowledge in multiple ways—formal or informal, structured or unstructured—a strong reason that speaks to the limitations of assuming uniformity in learning progressions.

Observation Model

The observation model concerns observations made from student responses to tasks and the assessment situation relevant to the target construct. It includes key features of assessment tasks that might elicit evidence related to that construct. The connections between the task and the target construct are established through identifying cognitive requirements for completing the tasks and describing how these requirements are related to the target construct and construct components in the cognition model.

In addition to how students learn, a myriad of factors that constitute the sociocultural context can affect how students relate to an assessment. Examples of these factors are how familiar students are with the items' contexts (the stories and situations provided by items with the intent to make problems meaningful; Ruiz-Primo & Li, 2015), how able they are to understand the items and communicate their answers, and how effectively they function and navigate in tests administered on digital platforms (Figure 3.5). In a science assessment, student perceptions of what science items are about, what they believe they are expected to do in responding to them, and what problem solution strategies they use to solve the problems may be shaped by their sociocultural activity (Solano-Flores & Nelson-Barber, 2001). In addition to these cognition-related factors, affective factors such as how students feel about the assessment situation, how they relate to the assessment tasks, and how their motivation is affected by the assessment contexts can impact how they approach and perform on an assessment (Pennington et al., 2016; Randall, 2021; Steele & Aronson, 1995).

This complex interaction between sociocultural context, student knowledge, and assessment challenges the assumption that responses to tasks can be interpreted as *equivalent evidence* of the construct for students from different sociocultural backgrounds and has serious implications concerning comparability and the use and meaningful interpretations of scores (Sireci, 2020). A key consideration is that the same task may impose different sets of cognitive demands that are not germane to the construct

**FIGURE 3.5**

Factors and Considerations Impacting the Observation Model

being assessed on students from different sociocultural backgrounds. Another key consideration is that students from different sociocultural backgrounds should not be expected to relate to that task in the same ways.

Grounding assessment in students' sociocultural contexts requires designing variations and features of tasks in ways that allow obtaining equivalent evidence of knowledge, skills, and competencies of students from different sociocultural backgrounds. We see examples of attempts to address such variations when tasks, task situations, or evaluation procedures are adapted to individual student contexts so that the resulting evidence is more comparable across individuals in tests adapted to different languages or modified to provide accommodations for specific student needs. It is important to note that because of cultural differences in practices and expectations regarding a construct, variations in tasks and task situations may not necessarily result in adequate comparability of evidence of the targeted constructs. An example of these types of large differences can be observed in cross-cultural intelligence testing (van de Vijver, 2003). Even though it may be possible to conceive of an "intelligence" construct with comparable elements across cultures, such capabilities are applied in extremely different activities in different cultures, which precludes the creation of tasks that will result in equivalent evidence and creation of a common assessment.

Two interrelated aspects of tasks are important to consider. One is the features of tasks intended to be related to providing construct-relevant evidence. The other feature

is the format and forms of representing information (e.g., wording, illustrations, and contextual information) that are not central to the target construct yet are needed to administer the task, but add cognitive demands irrelevant to the target construct. This is the case of text with directions explaining what the test taker is expected to do in a given task. Unless reading and interpreting the directions is the target construct, these directions inevitably add cognitive demands for completing the task.

In addition to these added cognitive demands, there is evidence that students from different cultural or linguistic backgrounds interpret graphics and images in tests differently (Solano-Flores et al., 2013), have different intrinsic interests in topics that impact their performance on tests (Ercikan, 1998), and use testing time differently (Ercikan et al., 2020). These findings underscore the key role that different semiotic (representational) resources play in designing tests and minimizing construct-irrelevant variance (Solano-Flores, 2021).

Fairness issues arise when these construct-irrelevant factors make assessment tasks systematically more challenging for students from certain backgrounds. In fact, the traditional notion of fairness is tied to the extent to which construct-irrelevant variance exists on a task that might be attributable to such presentation-related features. Minimizing construct-irrelevant variance involves the recognition of variations in cultural backgrounds, identities, reading ability, familiarity with context, and digital literacy among students and making choices about the ways in which tasks are to be presented.

In designing construct-relevant aspects of tasks, taking sociocultural context into account must also include considerations of the variations in hypothesized learning progressions and knowledge structures required to perform well on tasks across student groups. Unfortunately, while some researchers have recognized the influence of culture on students' learning progressions (e.g., Wilson & Toyama, 2018), empirical research has yet to systematically examine how these progressions vary across students from different cultural or linguistic groups. If learning progressions are substantially different across students from different backgrounds, then the features of the tasks that determine difficulty levels should not be assumed to be uniform across groups of students from different sociocultural backgrounds. Equity needs to be explicitly incorporated into the task of developing learning trajectories and learning progressions (Delgado & Morton, 2012).

The goal in designing both the presentation-related and the construct-related features of tasks from a sociocultural perspective is to contribute to obtaining valuable, equivalent evidence of knowledge, skills, and competencies for students from different sociocultural backgrounds. Consistent with Figure 3.5, possible guiding questions for developing and evaluating observation models include the following:

Presentation-related features:

- Are students from different backgrounds expected to relate to assessment tasks and the assessment situation differently?
- Are variations in sociocultural context expected to lead to different cognitive processes for students from different backgrounds?

- Are familiarity and working understanding of features used in a test (e.g., text, images, computer user interface) different across students from different backgrounds?
- Do these features need to be adapted to optimize performance across students from different sociocultural backgrounds?

Construct-related features:

- Are the definitions of the target construct different for different groups?
- Is there evidence of differences in learning progressions across groups?
- Is there evidence of differences in knowledge structures across groups?

Answers to these questions require empirical evidence, which typically is not available for all constructs that assessments may target. However, even in the absence of that evidence, and with input from experts in relevant sociocultural contexts, the questions should be used to guide explorations about the optimal design of tasks.

Interpretation Model

The interpretation model describes rationales and reasoning that are applied to interpreting observations (i.e., student responses, actions, products) as evidence of the target construct. It consists of two components, the *scoring model* and the *measurement model*. The scoring model, on the one hand, describes the essential features of the student responses that are relevant to the competencies in the cognition and learning model and assigns scores (ordered or partially ordered) or categorization codes to student responses or products. The scoring model is a characterization of the evidence in the performance that is relevant to making inferences about the target constructs. An example that illustrates a scoring model is the set of rules used to score constructed-response items or essays as indicators of performance in relation to the target construct. Other examples include evaluation and scoring of patterns of actions in a simulation-based task or an interactive communication assessment based on the degree to which different sequences of actions may be indicators of the targeted constructs at different levels.

The measurement model, on the other hand, is the psychometric approach for integrating performance data into evidence across student responses and products. For example, in the simplest case, using classical test theory, dichotomous scores on the correctness of the responses given by a student to multiple items can be added to compute a total score. Other measurement models include item response theory (IRT)–based models that define a probabilistic relationship between the performance and the target construct, Bayes nets, or integrated hidden Markov models that combine evidence from both responses and process data (Scalise et al., 2022). Measurement models facilitate the interpretation of observations in relation to what they indicate about the target constructs and involve making generalizations across individuals and contexts. Central to these interpretation models are standardized procedures for collecting data through tests and surveys to allow for the application of a single interpretation model across contexts. This standardization has utmost importance in creating comparable scores

and making defensible, consistent, and fair inferences about test takers' knowledge and skills (Moss et al., 2008).

Traditional interpretation models are silent about the sociocultural contexts of assessment and learning (see Figure 3.6). Typically, the same interpretation model is applied to individuals regardless of social, cultural, and language contexts. However, there is evidence that individuals engage in activities and tasks differently in different contexts (e.g., Beach, 1999; Lave, 1988; Mehan, 1998; Schratz & Mehan, 1993). For example, Lave (1988) demonstrated that some individuals who do not perform well on mathematics problem-solving tasks in formal testing situations were able to solve similar mathematical problems in everyday situations, such as determining the best supermarket buys. Lave argues that it is not possible to separate mathematical challenges from the sociocultural contexts and specific circumstances in which they are presented. These types of variations of performance across contexts limit the possibility of making generalizable claims about individuals' abilities.

Sociocultural context may impact student engagement with tasks in complex ways. Students' responses to tasks may vary because of differences in opportunity to learn given content and variations in access to resources that support learning. For example, the meaning and interpretation of the performance of students on active items that allow or require the use of calculators in a mathematics assessment are shaped by the kind of access students have to calculators at home and school. Imagine including calculators as a resource students can use optionally with the intent to eliminate calculation as an extraneous source of variance when calculation is not part of the target construct. This resource may make certain items easier for those students who are used to using calculators, but it may make them function differentially for those who are not accustomed to using calculators at home.

As with many tests being administered on digital platforms, access to technology and familiarity with different devices may impact performance on tests in different ways. There is extensive evidence of a digital divide among students from different cultural and socioeconomic backgrounds (e.g., Ching et al., 2005; Ercikan et al., 2018). There is also evidence that students' performance may be differentially impacted by the mode in which assessment is administered (e.g., paper based versus computer based) for students from different ethnic/racial and socioeconomic backgrounds (Jewsbury et al., 2017). These examples point to limitations of uniform interpretations of performance across groups. Therefore, in these circumstances, applying the same interpretation model might lead to inappropriate interpretations of response data and, therefore, to measurement error and bias.

Taking sociocultural context into account in the interpretation model requires considering the possibility that the same scoring model may not result in equivalent evidence of the target construct and that the same measurement model may not produce scores with equivalent meaning and, therefore, cannot be interpreted in the same way across students from different sociocultural contexts. Consistent with Figure 3.6, possible questions that need to be explored to examine the degree to which the same scoring and measurement models are expected to lead to valid interpretation and use of scores include the following:

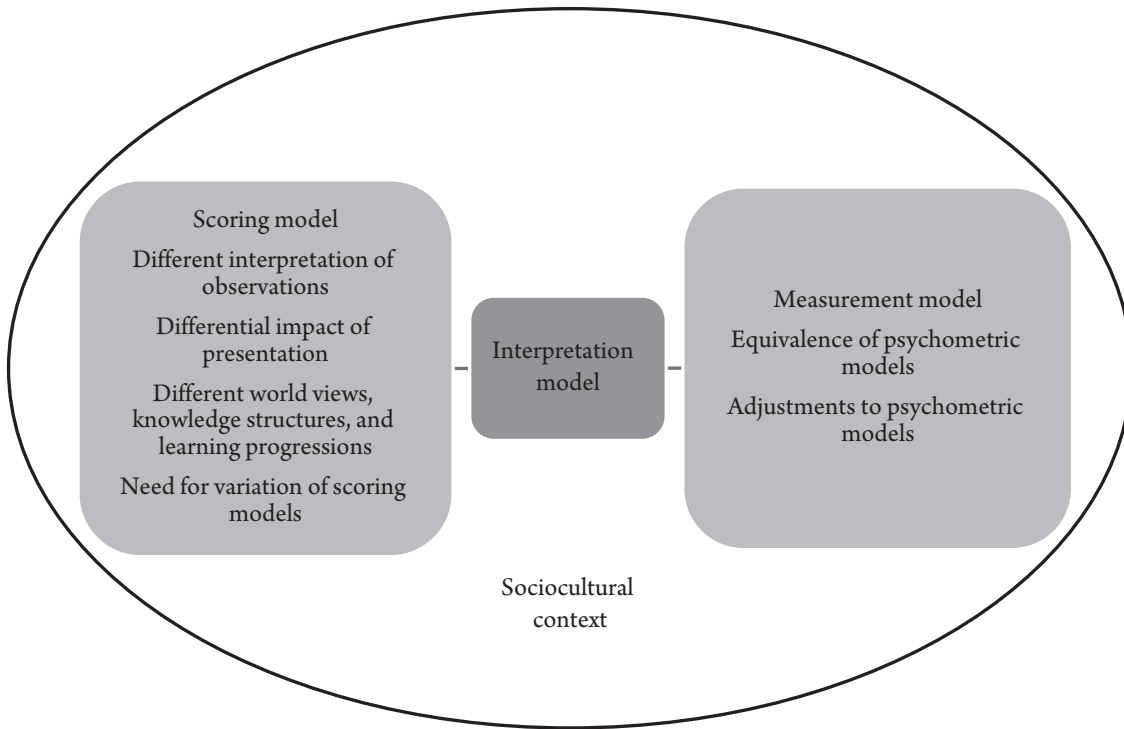


FIGURE 3.6
Factors and Considerations Impacting the Interpretation Model

Scoring model:

- How might the observations be interpreted differently in different sociocultural contexts?
- How might the presentation of items impact student performance differentially?
- How might different worldviews, knowledge structures, and learning progressions impact the nature and quality of observations?
- What types of variations in scoring models are needed for obtaining equivalent quality of evidence of the constructs from different student groups?

Measurement model:

- To what extent can the same psychometric model be used to summarize evidence across tasks for student groups from different sociocultural contexts?
- What adjustments might be needed in psychometric models for summarizing evidence of the construct with equivalent score meaning?

Effectively answering these questions requires developing scoring and psychometric models that recognize variations across different sociocultural contexts, as well as methodologies for examining the degree to which uniform interpretations of observations and accumulation of evidence are supported by response patterns. The considerations for all three models are summarized in Figure 3.7.

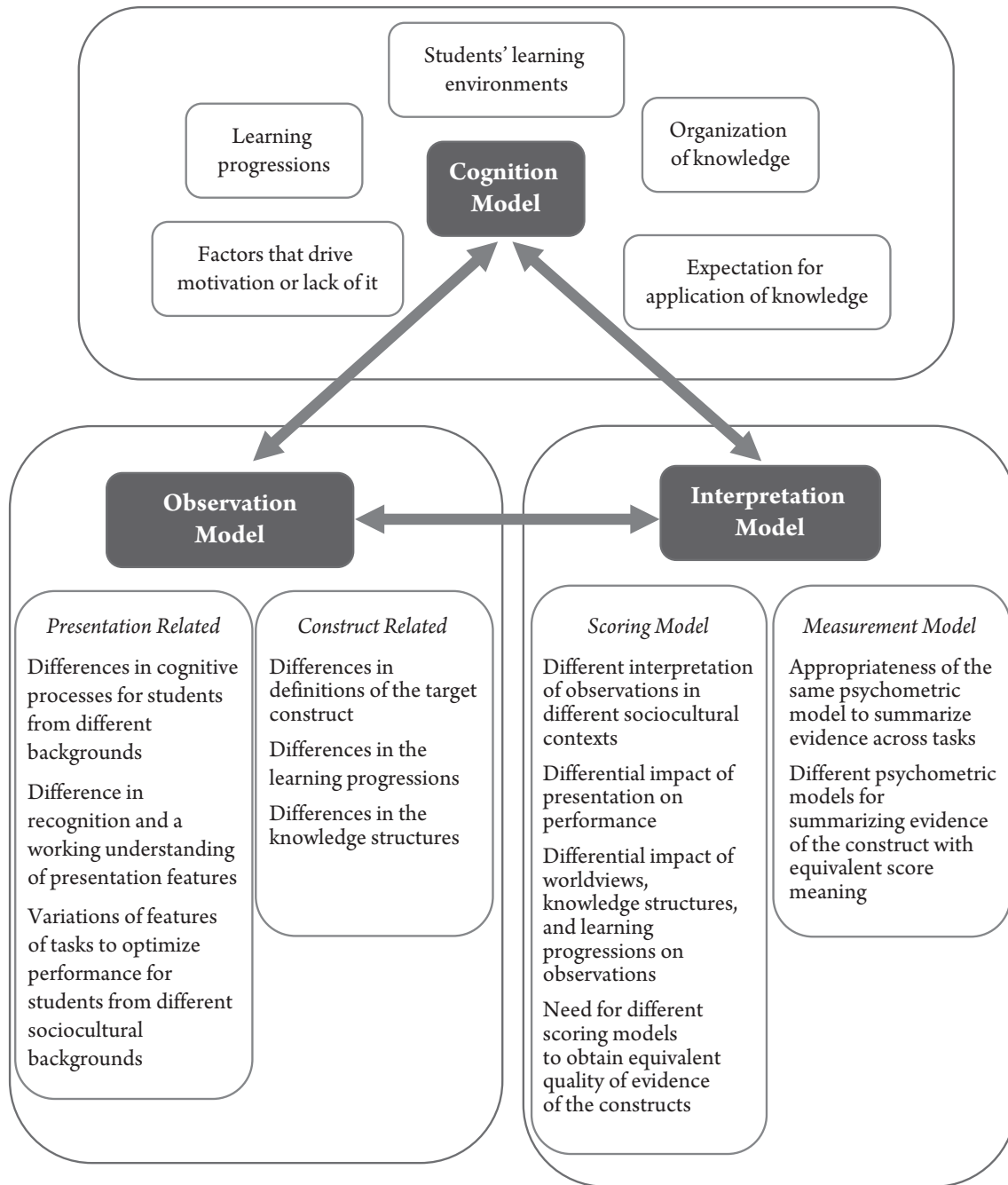


FIGURE 3.7
Sociocultural Considerations in the Assessment Triangle

IMPLICATIONS FOR FAIRNESS

We hope that in previous sections we have clearly argued that sociocultural context is relevant to all aspects of assessment, including how constructs are defined and developed, how performance on tasks can be designed, how measurement information can be derived, and how overall performance can be evaluated. This context is also central to the interpretation and use of assessment results, particularly to the consistency of such use and interpretation for individuals from different sociocultural contexts, which is at the core of fairness.

Two notions are central to Kane's (2006) view of fairness: (a) consistency of score meaning for groups and (b) equitable treatment of groups. *Consistency of score meaning* refers to the degree to which the same claims and inferences can be made based on test results for different groups. This notion of fairness has been the focus of much of fairness research. *Equitable treatment of groups* refers to the degree to which testing conditions are equivalent across groups. The notion of fairness as equitable treatment of groups has traditionally been addressed by standardization of testing conditions. The following sections discuss the sociocultural context and these two concepts of fairness.

Consistency of Score Meaning

To a great extent, measurement involves extracting and synthesizing evidence from students' performance on assessment tasks. A single measurement model is intended to summarize the relationship between observations and the target constructs. The measurement models used in synthesizing evidence regarding target constructs across tasks are intended to be used to make claims about all individuals. The appropriateness of the model as a method for summarizing evidence, as indicated by the fit of the model to the response patterns, is expected to vary across individuals and groups. Consistency of score meaning critically depends on the degree to which the same measurement adequately represents the relationship between response patterns and the target construct for different groups. Much of fairness research has been restricted to comparing the degree to which a given measurement model fits the response pattern data for individuals from different backgrounds—in other words, the degree to which a single measurement model can be justified for different groups.

There is evidence that speaks to the limited generalizability of models appropriate for overall groups to subgroups and individuals in the society (Ercikan & Roth, 2009). Social science research that relies on between-group score variations and marginal inference implicitly assumes generalizability of fundamental cognitive and affective processes across populations. However, there is a great degree of evidence of variations among human populations in visual perception, analytic reasoning, fairness, cooperation, and memory (Heinrich et al., 2010). Heinrich et al. (2010) highlighted that people from *Western, educated, industrialized, rich, and democratic* (WEIRD) societies are

among the most studied but are not necessarily the most typical. Findings from research that focuses on these restricted segments of societies hardly ever acknowledge the limitations that stem from the generalizability of its findings (Ercikan & Roth, 2009).

There are parallels in measurement. In North American contexts, with White, English-speaking mainstream cultures, estimation of measurement models through calibration analyses using IRT-based methodologies are heavily determined using response patterns from the mainstream culture through their higher degree of proportional representation in the calibration samples.

Measurement concepts and challenges to generalizability of measurement models are pertinent when taking sociocultural context into account. Because individuals from different sociocultural contexts and with different sets of learning experiences are likely to have differential response patterns, a single measurement model may not be equally effective in capturing the relationships between observations and the target construct for different groups.

Variations among contexts and individuals can result in differential response patterns and, therefore, in different levels of fit between the data and the model. These variations are captured by several measurement concepts. These include *person fit* (an indicator of the fit with the model summarizing relations among the tasks and the target construct for the individuals), *person-by-task interaction* (atypical response patterns for groups across tasks), and *group-by-task interaction*, also known as DIF. Researchers have highlighted the limitations of a single measurement model to *all groups* and *all individuals* (Ercikan et al., 2014; Molenaar, 1999, 2004; Molenaar et al., 2003; Oliveri et al., 2013, 2014). Borsboom et al. (2005) stated that “if a model fits in a given population, this does not entail the fit of the same model for any given element from a population, or even for the majority of elements from that population” (p. 73).

Person fit statistics allow the identification of individuals with different or atypical response patterns. Among other factors, poor person fit statistics could be indicators of confusion about the item’s content or presentation, lack of opportunities to learn the item’s cognitive requirements, and familiarity or lack thereof with the context in which the item is presented, which may impact the engagement of an individual with the item as compared to other persons. Inaccurate measurement results from estimating target competencies with a measurement model that has poor person fit. Appropriate uses of assessment results are critically dependent on the accuracy of measurement. The notion of accuracy as captured by person fit is different than reliability, which is a function of the correlation among the item responses calculated at the group level, or IRT-based measurement accuracy, based on information function values, also estimated at the group level. It is possible to have high reliability and information function values and still have poor person fit for some individuals and, therefore, low accuracy of score estimates. Typically, individuals from mainstream cultures tend to highly influence the overall estimation model. Therefore, individuals from nonmainstream cultures tend to be the most affected by poor person fit and, therefore, by inaccuracies of measurement. Different statistics developed for person fit (Meijer & Sijtsma, 2001; Rupp, 2013) are the best approaches created to address fairness for individuals.

Person-by-task interactions can be captured in generalizability theory models to identify the degree to which performance can be generalized across tasks to make claims about individuals' performances (see Cronbach et al., 1972). Large amounts of variance due to person-by-task interaction are common, which points to limitations in combining evidence of constructs across different types of tasks, such as hands-on tasks and multiple-choice questions. A sizable person-by-task interaction indicates variation of performance across both different types of tasks and individuals. It poses a limit to the use of a single model of synthesizing evidence across tasks (Shavelson et al., 1993). This interaction can be large when different constructs are measured by different tasks—in particular, by different task types. Large person-by-task interaction can also be an indication that fairness is compromised if its magnitude varies considerably across groups (Solano-Flores, 2017). This fact underscores the need to examine the technical properties of tests separately by group (e.g., the group-based variance components), which may reveal low technical properties of tests for some groups. It also speaks to the fact that generalizations based on a given test about specific individuals or about populations are tremendously dependent on the specific set of items included in the instrument and the specific conditions under which tests are administered.

Group-by-task interactions indicate systematic variations of group performance on specific tasks. These interactions are typically referred to as DIF and are central to examining consistency of score meaning for groups. These approaches play a critical role in identifying differential response patterns for *manifest* or observed groups identified as the focus of fairness research. However, a focus on sociocultural context and population heterogeneity highlights the limitations of these methodologies in examining fairness and identifying bias.

Population Heterogeneity and Fairness

IRT-based approaches to examining fairness are based on examining how differently items function across reference groups (typically, mainstream groups such as White/Caucasian or male groups) and focal groups (typically, historically underrepresented or underserved groups such as ethnic or linguistic minority groups). IRT-based DIF analysis is frequently invoked as a tool for examining bias and ensuring fairness in testing. However, DIF analysis methodologies are silent about the heterogeneity of populations. The configuration of focal groups often neglects the complexity of sociocultural contexts. This omission can lead to differential response patterns within such groups because of linguistic, cultural, and ethnic diversity (Cohen & Bolt, 2005; De Ayala et al., 2002; Ercikan, 2009).

Ignoring heterogeneity in comparison groups in DIF analyses compromises the fairness and consistency of score meaning. In DIF analyses, there are two possibilities: An item is identified as either differentially functioning or DIF free. When DIF detection involves heterogeneous group comparison groups, not all members of the focal group may be affected by DIF uniformly. Heterogeneity in response patterns may mask systematic differential response patterns between some subgroups within comparison groups and lead to false DIF-free identifications for some group comparisons. This can lead to inappropriate conclusions about item equivalence and score meaning for different groups.

The implication of heterogeneity within comparison groups is that items identified as differentially functioning and, therefore, potentially biased for the overall group (e.g., males or females) may not be functioning differentially for some of the subgroups within the larger group. Conversely, items that are identified as DIF free for the overall group may be differentially functioning for some subgroups. In essence, population heterogeneity may lead to inappropriate inferences in fairness research and consequently result in inconsistency of score meaning for some groups.

Researchers have recognized the possible limitations of DIF methodology as a result of diversity within comparison groups. Also, some researchers have highlighted heterogeneity within gender and ethnic groups and proposed examining possible interaction between these groupings (Dorans & Holland, 1993; Zhang et al., 2005). These researchers propose the creation of more homogeneous groups by crossing gender and ethnicity. While such an approach may indeed reduce heterogeneity within groups, it continues to rely on manifest grouping. Yet in the complex sociocultural context, the sources of variation in response patterns may not be easily identifiable through manifest labeling such as gender, ethnicity, socioeconomic status, or educational opportunity.

Other researchers have proposed consideration of latent groups that may be identifiable using statistical methods such as latent class modeling in fairness research. These researchers have proposed mixture model approaches for investigating differential response patterns (Bolt et al., 2001; Cohen & Bolt, 2005; De Ayala et al., 2002; Webb et al., 2008). In these models, also referred to as latent DIF models, comparisons focus on latent classes instead of manifest groupings. In mixture model approaches, the first step is to identify latent classes. Test takers are assigned to latent classes based on their item response patterns. Whereas manifest group-based DIF investigations typically focus on predetermined and potentially disadvantaged groups, this type of classification results in groupings that may have a combination of test takers from different ethnic or gender groups. One challenge with the mixture model approach to DIF is related to the interpretation of latent class membership for guiding fairness investigations.

Ercikan and Oliveri (2013) proposed an approach in which, first, heterogeneity within manifest groups is identified using latent class analysis and, second, DIF analysis for latent classes is performed. This two-step approach, which focuses on manifest groupings and addresses heterogeneity within these manifest groups, produces more accurate identifications of DIF and actionable steps following DIF investigations (Grover & Ercikan, 2017).

Equitable Treatment of Groups

Standardization has been central to the notion of fairness through the equitable treatment of groups, which assumes that even though some aspects in an assessment advantage some students, other aspects advantage others, and thus, on average, advantage/disadvantage evens out (Mislevy et al., 2013). However, as discussed before, standardization does not necessarily create equivalent testing conditions for individuals from different sociocultural backgrounds and, therefore, does not always provide equivalent evidence of students' knowledge, skills, and abilities. Individuals

with different experiences with testing (such as familiarity with testing formats and testing in technology-rich environments) and people from cultures and contexts who are not used to performing cognitive tasks in limited times or who relate to assessment tasks differentially may engage with tests differently. The extent to which testing conditions are equivalent has been investigated indirectly by comparing the performance of students from different ethnic and racial backgrounds on multiple-choice and performance tests (Ercikan et al., 1998), on paper-based and digital tests (Bennett et al., 2008), and across individuals from different sociocultural backgrounds (Ercikan et al., 2020). Despite pointing to the potential for nonequivalence of testing conditions, these approaches typically do not result in differential interpretations of scores.

Building on the notion of “marginal fairness” introduced by Hansen et al. (2005) for accessibility of tests, Mislevy et al. (2013) introduced the “conditional sense of fairness,” which

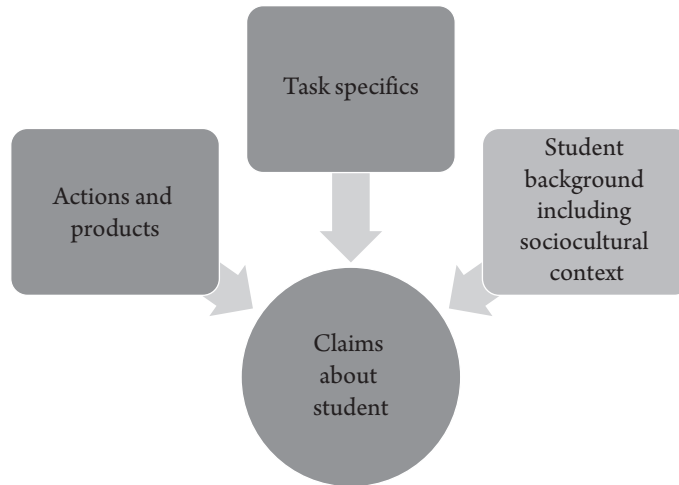
means taking certain information into account in inference rather than averaging over the ways it might vary. In assessment, it means a designer purposefully varies aspects of an assessment so that students can each access, interact with, and provide responses to tasks with minimal difficulty, and primary challenges are central to the intended construct. (Mislevy, 2018, p. 220)

This approach deviates from standardization and focuses on systematic variations of assessment and assessment conditions to provide equivalent evidence of the target construct. Examples of such deviations are observed on testing students with disabilities using accommodations of the tests (Kopriva et al., 2016; Mislevy et al., 2013; Thurlow & Kopriva, 2015) and adaptations of tests to different languages (Ercikan & Por, 2020; Solano-Flores et al., 2009).

Conditional Sense of Fairness and Measurement Models

The implementation of a conditional sense of fairness requires building conditional inference into assessment design and measurement models. Examples of conditional inference are presented by Mislevy et al. (2013), who proposed that tasks be matched to student populations. These researchers proposed integration of the universal design for learning with the evidence-centered design approach to assessment and drawing inferences from tests. The goal of this approach is to provide optimal conditions for students to demonstrate their knowledge, skills, and abilities in assessment situations. For example, students may be presented with different forms of the same task (e.g., braille versus regular text or different language versions) or they may be allowed to respond in different ways (e.g., in writing or orally).

Three sources of data are central to these inferential models: (a) student actions and products, (b) task specifics and contexts, and (c) student background data. Conditional inferences regarding student knowledge and abilities are conditional not only on student actions and products, but also on student background and task specifics (Figure 3.8). In this approach, the conditional relationship between responses, task context, and student background can be expressed in psychometric models for tests with multiple formats, such as those with accommodations presented to students with disabilities.

**FIGURE 3.8****Conditional Inference About Student Knowledge and Abilities**

Note. Adapted from “A ‘conditional’ sense of fairness in assessment,” by R. J. Mislevy, G. Haertel, B. H. Cheng, L. Rucitinger, A. DeBarger, E. Murray, and T. Vendliski, 2013, *Educational Research and Evaluation*, 19(2–3), p. 125. <https://doi.org/10.1080/13803611.2013.767614>.

Mislevy et al. (2013) discussed psychometric modeling of conditional inference when tasks are matched to students a priori. This modeling is proposed when students have varying levels of knowledge and skills that are irrelevant to the target construct yet are needed for reading, interpreting, and responding to tasks. Such a variation in construct-irrelevant skills can be observed for students with disabilities and ELs. Conditional inference is achieved by design by matching students to the appropriate task variations. This approach can be easily adopted when assessing students from diverse sociocultural contexts, provided that there are identifiable background variables of focus. These may include development of multiple versions of tasks, with different technology requirements for students with differential digital literacy or technology skills.

There are multiple complexities in identifying relevant sociocultural context factors, such as language, race/ethnicity, gender, intersectionality, and the features of tests that optimize performance for students from such backgrounds. Grounding assessments in sociocultural context requires gaining insights about aspects of sociocultural contexts that are of utmost importance to be reflected in an assessment situation, developing assessment task design strategies to reflect those aspects and optimize performance of students from all backgrounds, and developing measurement models that support making claims about diverse student populations based on such complex assessments. To do so requires a theory-based and empirically supported definition of the targeted construct as it relates to the intended interpretations and uses, the features of tasks and situations in which the construct is relevant, and the variety of necessary but construct-irrelevant features of situations that can be required to act in situations that can produce evidence of the intended construct.

FINAL NOTE

Our chapter is intended to help guide full consideration of the complex relationship of sociocultural context and assessment. Our main goal for the chapter is to raise awareness among measurement professionals, researchers, and practitioners of the importance of sociocultural context on the assessment, use, and interpretation of assessment results. We make recommendations for steps and considerations for a thoughtful approach to assessment. Such awareness is central to gauging how different contexts are affecting choices and decisions made in measurement, as well as in understanding the complex impact measurement has on different populations.

Our discussion has been grounded in the idea that social processes are central to learning and cognition and that individuals learn through social interaction, through learning from each other, and through motivating each other. In addition, sociocultural contexts influenced by adversity, poverty, and racism can have long-lasting impacts on health, well-being, learning, and achievement. From this sociocultural perspective, context shapes the mind, how knowledge is organized, how individuals approach solving problems, and how they engage with tests. Using the assessment triangle framework, we presented a set of considerations and issues to be addressed in each component of the triangle. These considerations are organized around questions intended to inform the definitions of each of the models and modify them for different sociocultural contexts.

Modifying each of these models can be challenging. We find some parallels when tests are modified to fit the needs of students with disabilities and ELs. However, these examples do not completely meet the requirements for tests in complex multicultural contexts. In the case of assessing students with disabilities and ELs, accommodations are made to the construct-irrelevant surface features of tests to optimize students' participation in the assessment. Therefore, only the observation and interpretation models are modified. In multicultural contexts, with evidence of different knowledge structures, modifications of the cognition model may also be needed. In that scenario, designing tests for variations of the cognition model will inevitably lead to variations of observation and interpretation models.

We offer two related approaches to addressing sociocultural context in assessment argumentation. One is the concept of *conditional inference*, which focuses on obtaining "equivalent evidence of the target construct" rather than standardized testing conditions for fairness of interpretations and uses of assessment results. The conditional inference approach allows for variations in observation and interpretation models. This means that variations in tasks for different sociocultural contexts are created to optimize testing conditions for different groups. It also means that responses can be interpreted differently using different scoring rubrics for extracting construct-relevant information from responses. We emphasize the recognition of population heterogeneity, moving away from the uniform treatment of groups such as language or cultural groups. This heterogeneity requires careful evaluations of the use of a single measurement model for different sociocultural contexts. Three sources of data are central to these measurement models: (a) student actions and products, (b) task specifics and contexts, and (c) student background data. Conditional

inference regarding student knowledge and abilities is conditional on student actions and products, but the student background and task specifics are relevant as well.

We believe that differences in the opportunities for learning the content assessed in tests, the cultural contexts in which students learn, and the variations among students in their ability to engage in assessment for students from different populations cannot be ignored. Validity and fairness are indissociable characteristics: One cannot address one while neglecting the other (Herman et al., 2023). Consideration of sociocultural contexts needs to occur continuously, from the conceptualization of a test and in all aspects of its development, interpretation, and use. Failure to take sociocultural context into account has significant consequences, including making inappropriate inferences about student learning, faulty decisions about individuals, unfair resource allocation and planning, and unequal distribution of opportunities to access education and careers.

Finally, this chapter should end by emphasizing the important role that organizations (e.g., local education agencies, assessment systems, test developers) play in ensuring that these ideas are applied to ensure fair and valid testing. While many of these methods and ideas have been available for decades, they are far from being consistently and systematically applied in practice. A serious commitment to fair, equitable testing necessarily entails a profound revision of assessment development procedures and a realistic allocation of the time and resources needed to properly address diversity. Clearly, the level of institutional commitment to fairness and equity is part of what defines the bidirectional relationship between sociocultural context and assessment.

ACKNOWLEDGMENTS

We thank the three reviewers, Bob Mislevy, Martha Thurlow, and Richard Duran, who reviewed and provided extensive feedback and directions in several iterations of this chapter. We are deeply grateful for their thoughtful suggestions and collegiality.

REFERENCES

- Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 277–398). Lawrence Erlbaum Associates.
- Abedi, J. (2012). Validity issues in designing accommodations for English language learners. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 62–76). Routledge.
- Abedi, J. (2014). English language learners with disabilities: Classification, assessment, and accommodation issues. *Journal of Applied Testing Technology*, 10(2), 1–30.
- Abedi, J. (2016). Language issues in item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 355–373). Routledge.
- Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium.
- Adewuya, A. O., & Ola, B. A. (2005). Prevalence of and risk factors for anxiety and depressive disorders in Nigerian adolescents with epilepsy. *Epilepsy & Behavior*, 6(3), 342–347.

- Albano, A. M., Chorpita, B. F., & Barlow, D. H. (2003). Childhood anxiety disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 279–329). Guilford Press.
- Alvero, A. J., Giebel, S., Gebre-Medhin, B., Antonio, A. L., Stevens, M. L., & Domingue, B. W. (2021). Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science Advances*, 7(42). <https://doi.org/10.1126/sciadv.abi9031>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Federation of Teachers. (2002). *Teaching English-language learners: What does research say?* (AFT Policy Brief Number 14). American With Disabilities Act. <https://adata.org/learn-about-ada>
- Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 et seq.
- Americans With Disabilities Act Amendments, 42 U.S.C. § 12101 et seq. (2008). <https://www.ada.gov/pubs/adastatute08.htm>
- Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing and student learning. *Education Policy Analysis Archives*, 10, 18. <https://doi.org/10.14507/epaa.v10n18.2002>
- Anderson, N. (2020, May 21). University of California takes huge step toward dropping SAT and ACT from admissions. *The Washington Post*. <https://www.washingtonpost.com/education/2020/05/21/sats-university-california-system/>
- Artiles, A. J., Kozleski, E. B., Trent, S. C., Osher, D., & Ortiz, A. (2010). Justifying and explaining disproportionality, 1968–2008: A critique of underlying views of culture. *Exceptional Children*, 76(3), 279–299.
- Bailey, A. L., & Durán, R. (2019). Language in practice: A mediator of valid interpretations of information generated by classroom assessments among linguistically and culturally diverse students. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 46–62). Routledge.
- Baker, B. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, 1052–1092. <https://doi.org/10.1007/s40593-021-00285-9>
- Baki, O., Erdogan, A., Kantarci, O., Akisik, G., Kayaalp, L., & Yalcikaya, C. (2004). Anxiety and depression in children with epilepsy and their mothers. *Epilepsy Behavior*, 5(6), 958–964.
- Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education*, 19, 115–132. https://doi.org/10.1207/s15324818ame1902_3
- Beach, K. (1999). Chapter 4: Consequential transitions: A sociocultural expedition beyond transfer in education. *Review of Research in Education*, 24(1), 101–139.
- Becker, K. A. (2003). *History of the Stanford–Binet intelligence scales: Content and psychometrics*. Riverside.
- Bejar, I. I. (2017a). A historical history of research regarding constructed-response formats. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 565–633). ETS.
- Bejar, I. I. (2017b). Threats to score meaning in automated scoring. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 75–84). Routledge.

- Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 28(2), 83–104. <https://doi.org/10.1080/10627197.2023.2202312>
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning and Assessment*, 6(9).
- Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvett, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10–11), 763–786.
- Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.). (2020). *Comparability of large-scale educational assessments: Issues and recommendations*. National Academy of Education.
- Bieber, T., Martens, K., Niemann, D., & Windzio, M. (2014). Grenzenlose bildungspolitik? Empirische evidenz für PISA als weltweites Leitbild für nationale Bildungsreformen. *Zeitschrift für Erziehungswissenschaft*, 17(S4), 141–166. <https://doi.org/10.1007/s11618-014-0513-6>
- Binet, A., & Simon, T. (1907). Le developpement de l'intelligence chez les enfants. *Année Psychologie*, 14, 1–94.
- Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). *School composition and the Black-White achievement gap* (NCES 2015-018). U.S. Department of Education, National Center for Education Statistics. <http://nces.ed.gov/pubsearch>.
- Bolt, D. M., Cohen, S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26(4), 381–409.
- Borrecia, C. P., Cheramie, G. M., & Borrecia, E. A. (2013). *Legal issues in educational testing*. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 517–542). American Psychological Association. <https://doi.org/10.1037/14049-025>
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Bourdieu, P. (1977). Cultural reproduction and social reproduction. In J. Karabel & A. H. Halsey (Eds.), *Power and ideology in education* (pp. 487–511). Oxford University Press.
- Braun, H. I., Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Borowiec, K. (2020). Performance assessment of critical thinking: Conceptualization, design, and implementation. *Frontiers in Education*, 5, 156. <https://doi.org/10.3389/educ.2020.00156>
- Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Paper Number 71). OECD Publishing.
- Bridgeman, B., Burton, N., & Cline, F. (2009). A note on presenting what predictive validity numbers mean. *Applied Measurement in Education*, 22(2), 109–119.
- Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1), 27–40.

- Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Taylor & Francis.
- Brown, G. T. L. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education*, 7, 1–8. <https://doi.org/10.3389/feduc.2022.1060633>
- Burstein, N., Cabello, B., & Hamann, J. (1993). Teacher preparation for culturally diverse urban students: Infusing competencies across the curriculum. *Teacher Education and Special Education*, 16(1), 1–13.
- Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 19(2–3), 104–120.
- Carnoy, M., & Rothstein, R. (2013). *What do international tests really show about U.S. student performance?* Economic Policy Institute. <https://www.issuelab.org/resources/14581/14581.pdf>
- Chavira, D. A., Stein, M. B., Bailey, K., & Stein, M. T. (2004). Child anxiety in primary care: Prevalent but untreated. *Depression and Anxiety*, 20(4), 155–164.
- Chi, M. T. H., Feltovich, P. G., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Ching, C. C., Basham, J. D., & Jang, E. (2005). The legacy of the digital divide: Gender, socioeconomic status, and early exposure as predictors of full-spectrum technology use among young adults. *Urban Education*, 40(4), 394–411.
- Civil Rights Act of 1964, Enrolled Acts and Resolutions of Congress, 1789–2011; General Records of the United States Government, Record Group 11; National Archives Building, Washington, DC.
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42, 133–148.
- Cole, M. (1999). Culture-free versus culture-based measures of cognition. In R. J. Sternberg (Ed.), *The nature of cognition* (pp. 645–664). MIT Press.
- Cole, M., Gay, J., Glick, J. A., & Sharp, D. W. (1971). *The cultural context of learning and thinking: An exploration in experimental anthropology*. Basic Books.
- Collins, J. (2009). Social reproduction in classrooms and schools. *Annual Review of Anthropology*, 38, 33–48. <https://doi.org/10.1146/annurev.anthro.37.081407.085242>
- Corbet, H. D., & Wilson, B. (1991). *Testing, rebellion and reform*. Ablex.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30(2), 116–127.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. Wiley.
- Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8(1), 1–41.
- De Ayala, R. J., Kim, S. H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2(3–4), 243–276.
- Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part*, 644 F.2d 397 (5th Cir. 1981); *on remand*, 564 F. Supp. 177 (M.D. Fla 1983), *aff'd*, 730 F.2d 1405 (11th Cir. 1984).

- Delgado, C., & Morton, K. (2012). Learning progressions, learning trajectories, and equity. In J. Van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012)–Volume 1, Full Papers* (pp. 204–211). International Society of the Learning Sciences.
- Department of Justice. (2010). *Americans with Disabilities Act Title III Regulations: Nondiscrimination on the Basis of Disability by Public Accommodations and in Commercial Facilities*. (March 8). <https://www.ada.gov/law-and-regs/regulations/title-iii-regulations/>
- DePascale, C., & Gong, B. (2020). *Comparability of individual students' scores on the "same test."*. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 25–48). National Academy of Education.
- DiCerbo, K. (2020). Assessment for learning with diverse learners in a digital world. *Educational Measurement: Issues and Practice*, 39(3), 90–93.
- Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: History, mathematics, and science in the classroom*. National Academies Press.
- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 1–40). Lawrence Erlbaum Associates.
- Dorsey, D. W., & Michaels, H. R. (2022). Validity arguments meet artificial intelligence in innovative educational assessment. *Journal of Educational Measurement*, 59(3), 267–271. <https://doi.org/10.1111/jedm.12331>
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287–301). Lawrence Erlbaum Associates.
- Drezner, D. W. (2001). Globalization and policy convergence. *International Studies Review*, 3(1), 53–78. <https://doi.org/10.1111/1521-9488.00225>
- Durán, R. P. (2008). Assessing English-language learners' achievement. *Review of Research in Education*, 32(1), 292–327. <https://doi.org/10.3102/0091732X07309372>
- Durán, R. P. (2011). Ensuring valid educational assessments for ELL students: Scores, score interpretation, and assessment uses. In M. R. Bastera, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 115–142). Routledge.
- ETS. (2022). *ETS guidelines for developing fair tests and communications*. <http://efaidnbmninnibpcajpcglclefindmkaj/https://www.ets.org/pdfs/about/fair-tests-and-communications.pdf>
- Elementary and Secondary Education Act of 1965, 89-10 U.S.C. § 1 (1965). <http://www.ed.gov/esea>
- Elman, B. A. (2013). The civil examination system in late imperial China, 1400–1900. *Frontiers of History in China*, 8(1), 32–50. <https://doi.org/10.3868/s020-002-013-0003-9>
- Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14(1), 133–156.

- Ercikan, K. (1998). Methodological issues in international assessments. *International Journal of Education*, 6(29), 487–489.
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage tests. *International Journal of Testing*, 2(3–4), 199–215.
- Ercikan, K. (2009). Limitations in sample to population generalizing. In K. Ercikan & M.-W. Roth (Eds.), *Generalizing in educational research: Beyond qualitative and quantitative polarization* (pp. 211–235). Routledge.
- Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, 29(2), 24–35.
- Ercikan, K., Asil, M., & Grover, R. (2018). Digital divide: A critical context for digitally based tests. *Education Policy Analysis Archives*, 26, 51. <https://doi.org/10.14507/epaa.26.3817>
- Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparison and fairness research. *Educational Assessment*, 25(3), 179–197. <https://doi.org/10.1080/10627197.2020.1804353>
- Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology* (Vol. 3, pp. 545–569). American Psychological Association.
- Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Educational Measurement*, 59(3), 272–287.
- Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity, fairness and testing of individuals in high stakes decision-making context* (pp. 69–86). Emerald Publishing.
- Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 1–8). Routledge.
- Ercikan, K., & Por, H. H. (2020). Comparability in multilingual and multicultural contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). National Academy of Education.
- Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, 35(5), 14–23.
- Ercikan, K., & Roth, W.-M. (Eds.). (2009). *Generalizing from educational research: Beyond qualitative and quantitative polarization*. Routledge.
- Ercikan, K., & Roth, W.-M. (2011). Constructing data. In C. Conrad & R. Serlin (Eds.), *The SAGE handbook for research in education* (2nd ed., pp. 219–245). Sage Publications.
- Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, 117(1), 1–28.

- Ercikan, K., Roth, M., Simon, M., Lyons-Thomas, J., & Sandilands, D. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, 27(4), 273–285.
- Ercikan, K., Schwarz, R., Julian, M., Burket, G., Weber, M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, 35(2), 137–155.
- Ercikan, K., & Solano-Flores, G. (2014). Introduction to the special issue: Levels of analysis in the assessment of linguistic minority students. *Applied Measurement in Education*, 27(4), 233–235.
- Ercikan, K., & Solano-Flores, G. (2016). Section discussion; Assessment and sociocultural context: A bidirectional relationship. In G. T. L. Brown & L. Harris (Eds.), *Handbook of human and social conditions of assessment* (pp. 490–505). Routledge.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (2nd ed.). MIT Press.
- Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). <https://www.ed.gov/essa?src=rn>
- Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g (1974). <https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html>
- Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmidt (Eds.), *Formal concept analysis: 4th International Conference, ICFCA 2006, Dresden, Germany, February 13–17, 2006 Proceedings, Intelligence* (pp. 61–79). Springer-Verlag.
- The Family Institute. (2024). *Americans With Disabilities Act*. <https://www.thearcfamilyinstitute.org/spceial-campiangs/ada.html>.
- Faulkner-Bond, M. M. (2016). *Who is like whom? Reclassification and performance patterns for different groupings of English learners* [Doctoral dissertation, University of Massachusetts Amherst]. https://scholarworks.umass.edu/dissertations_2/624
- Faulkner-Bond, M. (2020). Comparability when assessing English learner students. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 149–175). National Academy of Education.
- Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: An ironic investigation. *Journal of Education Policy*, 29(6), 845–855. <https://doi.org/10.1080/02680939.2014.892156>
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Educational Measurement*, 59(3), 288–313.
- Froese-Germain, B. (2010). *The OECD, PISA and the impacts on educational policy*. Canadian Teachers' Federation.
- Gándara, P., & Baca, G. (2008). NCLB and California's English language learners: The perfect storm. *Language Policy*, 7(3), 201–216.
- García, G. E., McKoon, G., & August, D. (2006). Language and literacy assessment of language-minority students. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 597–624). Lawrence Erlbaum Associates.

- Gebril, A. (2016). Educational assessment in Muslim countries: Values, policies, and practices. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 420–435). Routledge.
- Gee, J. P. (2008). Learning and games. In K. Salen & K. S. Tekinbas (Eds.), *The ecology of games: Connecting youth, games, and learning, The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning* (pp. 21–40). MIT Press.
- Gelman, S. A., & Roberts, S. O. (2017). How language shapes the cultural inheritance of categories. *Proceedings of the National Academy of Sciences*, 114(30), 7900–7907. <https://doi.org/10.1073/pnas.1621073114>
- General Accounting Office. (1993). *Student testing: Current extent and expenditures, with costs estimates for a national examination* [Report to Congressional Requesters].
- Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement*, 38(2), 164–187.
- Gilmour, A. F., Fuchs, D., & Wehby, J. H. (2019). Are students with disabilities accessing the curriculum? A meta-analysis of the reading achievement gap between students with and without disabilities. *Exceptional Children*, 85(3), 329–346. <https://doi.org/10.1177/0014402918795830>
- Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Prentice Hall.
- Goldman, S. R., & Lee, C. D. (2024). Human learning and development: Theoretical perspectives to inform assessment systems. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 48–92). National Academy of Education.
- Goldrick-Rab, S., Richardson, J., & Hernandez, A. (2017). *Hungry and homeless in college: Results from a national study of basic needs insecurity in higher education*. Wisconsin HOPE Lab.
- Gottlieb, M., Cranley, M. E., & Camilleri, A. (2007). *Understanding the WIDA English language proficiency standards: A resource guide*. Board of Regents of the University of Wisconsin System.
- Gould, S. J. (1981). The mismeasure of man. *Bridgewater Review*, 1(1), 12–13.
- Green, B. F. (1981). A primer of testing. *American Psychologist*, 36(10), 1001–1011. <https://doi.org/10.1037/0003-066X.36.10>
- Greenfield, P. M. (1997). Culture as process: Empirical methods for cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Vol. 1. Theory and method* (2nd ed., pp. 301–346). Allyn & Bacon.
- Greeno, J. G. (2002). *Students with competence, authority, and accountability: Affording intellectual identities in classrooms*. College Board.
- Greeno, J. G., & Gresalfi, M. S. (2008). Opportunities to learn in practice and identity. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 170–199). Cambridge University Press. <https://doi.org/10.1017/CBO9780511802157.009>
- Grek, S. (2012). What PISA knows and can do: Studying the role of national actors in the making of PISA. *European Educational Research Journal*, 11(2), 243–254. <https://doi.org/10.2304/eeerj.2012.11.2.243>

- Grover, R., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education*, 30(3), 178–195.
- Guo, H., & Ercikan, K. (2021). *Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: Use of response time in measurement comparability research* (ETS Research Report No. RR-21-25). ETS.
- Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. R. Gifford (Ed.), *Test policy and test performance: Evaluation in education, language, and culture* (pp. 25–50). Springer. https://doi.org/10.1007/978-94-009-2500-7_2
- Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics*, 33, 107–133.
- Hambleton, R. K. (2001). The generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, 17(3), 164–172.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Associates.
- Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science tests. *Applied Measurement in Education*, 10(2), 181–200.
- Haney, W. (1984). Testing reasoning and reasoning about testing. *Review of Educational Research*, 54(4), 597–654.
- Hatzenbuehler, M. L., & Link, B. G. (2014). Introduction to the special issue on structural stigma and health. *Social Science & Medicine*, 103, 1–6.
- Hedges, L. V. (1987). How hard is hard science and how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, 42(2), 443–455.
- Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29. <https://doi.org/10.1038/466029a>
- Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, 12(4), 20–25.
- Herman, J. L., Martínez, J. F., & Bailey, A. L. (2023). Fairness in educational assessment and the next edition of the *Standards*: Concluding commentary. *Educational Assessment*, 28(2), 128–136. <https://doi.org/10.1080/10627197.2023.2215980>
- Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 37–75). CRC Press.
- Hommel, K. A., Chaney, J. M., Wagner, J. L., White, M. M., Hoff, A. L., & Mullins, L. L. (2003). Anxiety and depression in older adolescents with long-standing asthma: The role of illness uncertainty. *Children's Health Care*, 32(1), 51–63.
- Hsieh, H., Wiles, S., & Solano-Flores, G. (2023). How different are English learners from their native English-speaking peers? Evidence of equivalent lexical competence in classroom conversations. *International Multilingual Research Journal*, 17(2), 87–106. <https://doi.org/10.1080/19313152.2022.2118193>
- Huberty, T. J. (2012). *Anxiety and depression in children and adolescents: Assessment, intervention, and prevention*. Springer.

- Hubbler, S. (2023, May 23). Why is SAT falling out of favor? *The New York Times*. <https://www.nytimes.com/2020/05/23/us/SAT-ACT-abolish-debate-california.html>
- Huebeck, E. (2024, February 28). Dartmouth and Yale are backtracking on “test-optional” admissions. Why that matters. *Education Week*. <https://www.edweek.org/teaching-learning/dartmouth-and-yale-are-backtracking-on-test-optional-admissions-why-that-matters/2024/02>
- Huebert, J. P., & Hauser, R. M. (Eds.). (1998). *High-stakes testing for tracking, promotion, and graduation*. National Academies Press.
- Hughes, G. B. (2023). *Improving equitable measurement and reporting in NAEP*. American Institutes for Research.
- Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries*. Wiley.
- Husén, T. (1983). *An incurable academic: Memoirs of a professor*. Pergamon Press.
- Individuals With Disabilities Education Improvement Act, 20 U.S.C. § 1400 (2004). <https://sites.ed.gov/idea/statuteregulations/>
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). <https://www.intestcom.org/page/14>
- Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, 43(8), 381–389.
- Jewsbury, P., Finnegan, R., Xi, N., Jia, Y., Rust, K., Burg, S., Donahue, P., Mazzeo, J., Cramer, B., Lin, A., & Weil, N. (2017). *NAEP transition to digitally based assessments in mathematics and reading at Grades 4 and 8: Mode evaluation study*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf
- Jonsson, A., & Leden, L. (2019). The ambiguous influence of high-stakes testing on science teaching in Sweden. *International Journal of Science Education*, 41(14), 1926–1943.
- Kachchaf, R. R. (2011). *Exploring problem solving strategies on multiple-choice science items: Comparing native Spanish-speaking English language learners and mainstream monolinguals* [Doctoral dissertation, University of Colorado Boulder]. https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/t435gd12s
- Kachchaf, R., Noble, T., Rosebery, A., O'Connor, C., Warren, B., & Wang, Y. (2016). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal*, 39(2), 152–166. <https://doi.org/10.1080/015235882.2016.1169455>
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9), 697–720.
- Kane, M. T. (1982). A sampling model for validity. *Applied Measurement in Education*, 6(2), 125–160. <https://doi.org/10.1177/014662168200600201>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education; Praeger.
- Kane, M. T. (2012, March 28–29). *Validity, fairness, and testing* [Paper presentation]. Educational Assessment, Accountability, and Equity: Conference on Conversations

- on Validity Around the World. Teachers College, New York, NY, United States. <https://www.tc.columbia.edu/media/media-library-2018/centers-amp-labs/aeri/conferences-and-forms-/conversations-on-validity-2012-/a283745b-70b8-486c-bdee-4e1f4cda2c48.pdf>
- Kane, T., & Reardon, S. (2023, May 11). Parents don't understand how far behind their kids are in school. *The New York Times*. <https://www.nytimes.com/interactive/2023/05/11/opinion/pandemic-learning-losses-steep-but-not-permanent.html>
- Katz, I. R., & Gorin, J. (2016). Computerising assessment: Impacts on education stakeholders. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 472–489). Routledge.
- Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian heritage cultures. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 407–419). Routledge.
- King, A. G., & Mrkonich, M. J. (2016). Big data and the risk of employment discrimination. *Oklahoma Law Review*, 68(3), 555–584.
- Klingner J., & Harry, B. (2006). The special education referral and decision-making process for English language learners: Child study team meetings and placement conferences. *Teachers College Record*, 108(11), 2247–2281.
- Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist*, 51(1), 108–128.
- Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education; Praeger.
- Kukea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education*, 6, 701973. <https://doi.org/10.3389/feduc.2021.701973>
- Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge University Press.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Leibert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, 20(3), 975–978.
- Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. *Applied Measurement in Education*, 26(2), 136–157. <https://doi.org/10.1080/08957347.2013.765435>
- Leighton, J. P. (2017). Collecting and analyzing verbal response process data in the service of interpretive and validity arguments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 25–38). Routledge.

- Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.
- Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement*, 20(2), 179–189.
- Linquanti, R., Cook, H. G., Bailey, A. L., & MacDonald, R. (2016). *Moving toward a more common definition of English learner: Collected guidance for states and multi-state assessment consortia*. Council of Chief State School Officers.
- Lipka, J., Andrew-Ihrke, D., Koester, D., Zinger, V., Olson, M., Yanez, E., & Rubinstein, D. (2015). Indigenous knowledge provides an elegant way to teach the foundations of mathematics. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 2–18). Michigan State University. <https://files.eric.ed.gov/fulltext/ED584219.pdf>
- MacMillan, D. L., & Hendrick, I. G. (1993). Evolution and legacies. In K. I. Goodlad & T. C. Lovitt (Eds.), *Integrating general and special education* (pp. 23–48). Merrill.
- Manyika, J., Silberg, J., & Presten, B. (2019, October 25). What do we do about the biases in AI? *Harvard Business Review*, 25. <https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai>
- Martinez, Y., & Ercikan, K. (2009). Chronic illnesses in Canadian children: What is the effect of illness on academic achievement, and anxiety and emotional disorders? *Child: Care, Health and Development*, 35(3), 391–401.
- Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333–368.
- Maul, A., Torres Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, 79, 311–320. <http://doi.org/10.1016/j.measurement.2015.11.001>
- Mehan, H. (1998). The study of social interaction in educational settings: Accomplishments and unresolved issues. *Human Development*, 41(4), 245–269.
- Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archive*, 6(13), 1–30. <https://epaa.asu.edu/index.php/epaa/article/view/580/703>
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.
- Meltzoff, A. N., Kuhl, P. K., Movellan, J., & Sejnowski, T. J. (2009). Foundations for a new science of learning. *Science*, 325(5938), 284–288.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Milner, A. R., Sondergeld, T. A., Demir, A., Johnson, C. C., & Czerniak, C. M. (2012). Elementary teachers' beliefs about teaching science and classroom practice: An examination of pre/post NCLB testing in science. *Journal of Science Teacher Education*, 23(2), 111–132. <https://doi.org/10.1007/s10972-011-9230-7>
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

- Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., & Vendlinski, T. (2013). A “conditional” sense of fairness in assessment. *Educational Research and Evaluation*, 19(2–3), 121–140. <https://doi.org/10.1080/13803611.2013.767614>
- Mlodinow, L. (2008). *The drunkard's walk: How randomness rules our lives*. Vintage Books/Random House.
- Molenaar, P. C. M. (1999). Longitudinal analysis. In H. J. Adèr & G. J. Mellenbergh (Eds.), *Research methodology in the life, behavioural and social sciences* (pp. 143–167). Sage.
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201–218.
- Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of inter-individual and intra-individual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development* (pp. 339–360). Kluwer Academic Publishers.
- Morsy, L., Khavenson, T., & Carnoy, M. (2018). How international tests fail to inform policy: The unsolved mystery of Australia's steady decline in PISA scores. *International Journal of Education and Development*, 60(C), 60–79. <https://doi.org/10.1016/j.ijedudev.2017.10.018>
- Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press.
- Mullainathan, S., & Shafir, E. (2013). *Scarcity: Why having too little means so much*. Macmillan.
- Nasir, N. S., de Royston, M. M., Barron, B., Bell, P., Pea, R., Stevens, R., & Goldman, S. (2020). Learning pathways: How learning is culturally organized. In N. S. Nasir, C. D. Lee, R. Pea, & M. M. de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 195–211). Routledge.
- Nasir, N. S., Lee, C. D., Pea, R., & de Royston, M. M. (Eds.) (2020). *Handbook of the cultural foundations of learning*. Routledge.
- National Academies of Sciences, Engineering, and Medicine. (2017). *Promoting the educational success of children and youth learning English: Promising futures*. The National Academies Press. <https://doi.org/10.1722/24677>
- National Academies of Sciences, Engineering, and Medicine. (2018a). *English learners in STEM subjects: Transforming classrooms, schools, and lives*. The National Academies Press. <https://doi.org/10.17226/25182>
- National Academies of Sciences, Engineering, and Medicine. (2018b). *How people learn II: Learners, contexts, and cultures*. National Academies Press. <https://fji.oer4pacific.org/id/eprint/13/1/NAP-BOOK.pdf>
- National Center for Homeless Education. (2022). Student homelessness in America: School years 2018–19 to 2020–21. University of North Carolina at Greensboro.

- National Research Council. (1997). *Educating one and all: Students with disabilities and standards-based reform*. The National Academies Press. <https://doi.org/10.17226/5788>.
- National Research Council. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academies Press.
- National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). National Academies Press.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessments*. National Academies Press.
- Nichols, P. D., Kobrin, J. J., Lai, E., & Koepfler, J. (2017). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 15–40). John Wiley & Sons.
- Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G. M., Grundy, S., Lyden, S., Neal, P., & Sandison, C. (2023). ChatGPT versus engineering education assessment: A multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. *European Journal of Engineering Education*, 48(4), 559–614. <https://doi.org/10.1080/03043797.2023.2213169>
- Noble, S. U. (2018). *Algorithms of oppression. How search engines reinforce racism*. New York University Press.
- No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2001). <https://www.congress.gov/bill/107th-congress/house-bill/1>
- Obradovic, J., & Armstrong-Carter, E. (2021). Addressing educational inequalities and promoting learning through studies of stress physiology in elementary school students. *Development and Psychopathology*, 32(5), 1899–1913.
- Office of Technology Assessment. (1992, February). *Testing in American schools: Asking the right questions* (Vol. 22, Report No. OTA-SET-519). U.S. Government Printing Office.
- Oliveri, M. E., Ercikan, K., & Simon, M. (2015). A framework for developing comparable multi-lingual assessments for minority populations: Why context matters. *International Journal of Testing*, 15(2), 94–113.
- Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class DIF in international assessments. *International Journal of Testing*, 13(3), 272–293.
- Oliveri, M. E., Ercikan, K., & Zumbo, B. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, 27(4), 286–300.
- Oranje, A., Gorin, J., Ha, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 39–51). Routledge.
- Overbaugh, R., & Lu, R. (2008). The impact of a NCLB–EETT funded professional development program on teacher self-efficacy and resultant implementation.

- Journal of Research on Technology in Education*, 41(1), 43–61. <https://doi.org/10.1080/15391523.2008.10782522>
- Owens, K., & Kaleva, W. (2007). Changing our perspective on measurement: A cultural case study. In J. Watson & K. Beswick (Eds.), *Mathematics: Essential research, essential practice—Proceedings of the 30th Annual Conference of the Mathematics Education Research Group of Australasia, Hobart* (pp. 571–580). MERGA. <https://files.eric.ed.gov/fulltext/ED503746.pdf>
- Pennington, C. R., Heim, D., Levy, A. R., & Larkin, D. T. (2016). Twenty years of stereotype threat research: A review of psychological mediators. *PloS One*, 11(1), e0146487. <https://doi.org/10.1371/journal.pone.0146487>
- Pepper, D., Hodgen, J., Lamesoo, K., Kõiv, P., & Tolboom, J. (2018). Think aloud: Using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics. *International Journal of Research & Method in Education*, 41(1), 3–16. <https://doi.org/10.1080/1743727X.2016.1238891>
- Personnel Administrator of Mass. v. Feeney, 442, U.S. 256, 99S. Ct. 2282, 0 L. Ed. 2d 870 (1979).
- Peterson, J. J., Lowe, J. B., Peterson, N. A., & Janz, K. F. (2006). The relationship between active living and health-related quality of life: Income as a moderator. *Health Education Research*, 21(1), 146–156. <https://doi.org/10.1093/her/cyh050>
- Phelps, R. P. (2000). Estimating the cost of standardized student testing in the United States. *Journal of Education Finance*, 25(3), 343–380.
- Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (pp. 733–755). American Council on Education; Praeger.
- Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16–20.
- Puffer, E. S., Giusto A., Rieder, A. D., Friis-Healy, E., Ayuku, D., & Green, E. P. (2021). Development of the family togetherness scale: A mixed-methods validation study in Kenya. *Frontiers in Psychology*, 12(662991). <https://doi.org/10.3389/fpsyg.2021.662991>
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.
- Randall, J. (2023). It ain’t near ‘bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment*, 28(2), 68–82. <https://doi.org/10.1080/10627197.2023.2223924>
- Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon (Ed.), *Handbook of child psychology: Vol. 2. Cognition, perception, and language* (pp. 679–744). John Wiley & Sons.
- Rogoff, B., & Lave, J. (Eds.). (1984). *Everyday cognition: Its development in sociocultural context*. Harvard University Press.
- Rogoff, B., Paradise, R., Arauz, R. M., Correa-Chávez, M., & Angelillo, C. (2003). Firsthand learning through intent participation. *Annual Review of Psychology*, 54(1), 175–203.

- Rose, D., Murray, E., & Gravel, J. (2012). *UDL and the PADI process: The foundation* (Assessment for Students With Disabilities Technical Report 4). SRI International.
- Roth, W.-M., & Lee, Y. J. (2007). "Vygotsky's neglected legacy": Cultural-historical activity theory. *Review of Educational Research*, 77(2), 186–232.
- Ruiz-Primo, M. A., & Furtak, E. M. (2024). Classroom activity systems to support ambitious teaching and assessment. In S. F. Marion, J. W. Pellegrino, & A. I. Berman (Eds.), *Reimagining balanced assessment systems* (pp. 93–131). National Academy of Education.
- Ruiz-Primo, M. A., & Li, M. (2015). The relationship between item context characteristics and student performance: The case of the 2006 and 2009 PISA science items. *Teachers College Record*, 117(1), 1–36.
- Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative mapping techniques. *Educational Assessment*, 7(2), 99–141.
- Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, 55, 3–38.
- Russell, M. (2023). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond*. Routledge.
- Scalise, K., Malcom, C., & Kaylor, E. (2022). Innovating assessments to measure and support complex skills. In N. Foster & M. Piacentini (Eds.), *Analysing and integrating new sources of data reliably in innovative assessments: Innovating assessments to measure and support complex skills* (pp. 190–210). OECD Publishing.
- Schratz, M., & Mehan, H. (1993). Gulliver travels into a math class. In search of alternative discourse in teaching and learning. *International Journal of Educational Research*, 19, 247–247.
- Section 504, Rehabilitation Act of 1973, 29 U.S.C. § 701 (1973). <https://www2.ed.gov/about/offices/list/ocr/docs/edlite-FAPE504.html#:~:text=The%20publication's%20citation%20should%20be,%2C%20Washington%2C%20D.C.%2C%202010>
- Setati, M., & Adler, J. (2001). Between languages and discourses: Language practices in primary multilingual mathematics classrooms in South Africa. *Educational Studies in Mathematics*, 43(3), 243–269.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance tests. *Journal of Educational Measurement*, 30(3), 215–232.
- Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan*, 71(9), 692–697.
- Shaw, S. R., & Páez, D. (2002). Best practices in interdisciplinary service delivery to children with chronic medical issues. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1473–1483). National Association of School Psychologists.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19(1), 405–450.

- Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, 31, 100335.
- Simon, M., Ercikan, K., & Rousseau, M. (Eds.). (2013). *Improving large-scale assessment in education: Theory, issues, and practice*. Routledge.
- Singer, J. D., Braun, H. I., & Chudowsky, N. (2018). *Methods and policy uses of international large-scale assessments: Report of a workshop*. National Academy of Education.
- Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, 16(1), 12–19.
- Sireci, S. G. (2020). Standardization and UNDERSTANDARDIZATION in educational assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105. <https://doi.org/10.1111/emip.12377>
- Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, 39(1), 215–252. <https://doi.org/10.3102/0091732X14557003>.
- Sireci, S., Scarpatti, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490. <https://journals.sagepub.com/doi/pdf/10.3102/00346543075004457>
- Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English-language learners. *Teachers College Record*, 108(11), 2354–2379.
- Solano-Flores, G. (2009). The testing of English language learners as a stochastic process. In K. Ercikan & W-M. Roth, (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 33–45). Routledge.
- Solano-Flores, G. (2012). *Smarter Balanced Assessment Consortium: Translation accommodations framework for testing English language learners in mathematics*. Smarter Balanced Assessment Consortium. <http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/09/Translation-Accommodations-Framework-for-Testing-ELL-Math.pdf>
- Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education*, 27(4), 236–247.
- Solano-Flores, G. (2017). Generalizability. In D. Wyse, N. Selwyn, E. Smith, & L. E. Suter (Eds.), *Handbook of educational research* (pp. 937–956). Sage. <http://doi.org/10.4135/9781473983953.n47>
- Solano-Flores, G. (2019). The participation of Latin American countries in international assessments: Assessment capacity, validity, and fairness. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *Sage handbook on comparative studies in education: Practices and experiences in student schooling and learning* (pp. 139–161). Sage.
- Solano-Flores, G. (2021). The semiotics of test design: Conceptual framework on optimal item features in educational assessment across cultures, countries, and languages. *Frontiers in Education* (Vol. 6, p. 637993). <https://doi.org/10.3389/feduc.2021.637993>
- Solano-Flores, G. (2022). Fairness in testing: Designing, using, and evaluating test accommodations for English learners. In J. L. Jonson & K. R. Geisinger (Eds.),

- Fairness in educational and psychological testing: Examining theoretical research, practice, and policy implications of the 2014 Standards* (pp. 271–292). American Educational Research Association.
- Solano-Flores, G. (2023). How serious are we about fairness in testing and how far are we willing to go? A response to Randall and Bennett with reflections about the Standards for Educational and Psychological Testing. *Educational Assessment*, 28(2), 105–117, <https://doi.org/10.1080/10627197.2023.2226388>
- Solano-Flores, G., Backhoff, E., & Contreras-Nino, L. A. (2009). Theory of test translation error. *International Journal of Testing*, 9(2), 78–91.
- Solano-Flores, G., Backhoff, E., Contreras-Niño, L. A., & Vázquez-Muñoz, M. (2015). Language shift and the inclusion of Indigenous populations in large-scale assessment programs. *International Journal of Testing*, 15(2), 136–152.
- Solano-Flores, G., Barnett-Clarke, C., & Kachchaf, R. (2013). Semiotic structure and meaning making: The performance of English language learners on mathematics tests. *Educational Evaluation*, 18(3), 147–161.
- Solano-Flores, G., & Hakuta, K. (2017, May 5). *Assessing students in their home language*. Understanding Language/Stanford University. <https://portal.smarterbalanced.org/library/en/assessing-students-in-their-home-language.pdf>
- Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, 28(2), 9–18.
- Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, 19(2–3), 245–263.
- Solano-Flores, G., & Milbourn, T. (2016). Assessment capacity, cultural validity and consequential validity in PISA. *Relieve*, 22(1), 1–16. <http://doi.org/10.7203/relieve.22.1.8281>
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
- Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, “translations” and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, 19(3), 387–420. <https://doi.org/10.1177/030631289019003001>
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797–811. <https://doi.org/10.1037/0022-3514.69.5.797>
- Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology*, 34, 379–440. <https://cpb-us-w2.wpmucdn.com/u.osu.edu/dist/2/43662/files/2017/02/Steele-Spencer-Aronson-2002.-Contending-with-group-image-1ji15ji.pdf>
- Takanishi, R., & Le Menestrel, S. (2017). *Promoting the educational success of children and youth learning English: Promising futures*. The National Academies Press.

- Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass*, 8(12), 1159–1173. <https://doi.org/10.1111/phc3.12089>
- Taras, H., & Potts-Datema, W. (2005). Childhood asthma and student performance at school. *Journal of School Health*, 75(8), 296–312.
- Taylor, C. S. (2022). *Culturally and socially responsible assessment: Theory, research, and practice*. Teachers College Press.
- Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 369–386). Routledge.
- Thurlow, M. L., & Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education*, 39(1), 331–369.
- Tomasello, M. (2009) *Why we cooperate*. MIT Press.
- Tong, Y. (2022). NCME presidential address 2021: Assessment research and practice in the post-COVID-19 era. *Educational Measurement: Issues and Practice*, 41(2), 7–11.
- Trumbull, E., & Nelson-Barber, S. (2019). The ongoing quest for culturally-responsive assessment for Indigenous students in the U.S. *Frontiers in Educations*, 4, 40. <https://doi.org/10.3389/feduc.2019.00040>
- Tyack, D. B. (1974). *The one best system: A history of American urban education* (Vol. 95). Harvard University Press.
- United States v. Fordice, Governor of Mississippi et al., 505 U.S. 717 (1992). <https://tile.loc.gov/storage-services/service/lh/usrep/usrep505/usrep505717/usrep505717.pdf>
- Valdés, G. (2004). Between support and marginalisation: The development of academic language in linguistic minority children. In J. Brutt-Griffler & M. Varghese (Eds.), *Bilingualism and language pedagogy* (pp. 10–41). Multilingual Matters.
- Valdés, G. (2018). Analyzing the curricularization of language in two-way immersion education: Restating two cautionary notes. *Bilingual Research Journal*, 41(4), 388–412. <https://doi.org/10.1080/15235882.2018.1539886>
- Valdés, G., & Figueroa, R. (1994). *Bilingualism and testing: A special case of bias*. Ablex.
- Van Ameringen, M., Mancini, C., & Farvolden, P. (2003). The impact of anxiety disorders on educational achievement. *Journal of Anxiety Disorders*, 17(5), 561–571.
- van de Vijver, F. J. R. (2003). Principles of adaptation of intelligence tests to other cultures. In J. Georgas, F. J. R. van de Vijver, L. G. Weiss, & D. H. Saklofske (Eds.), *Practical resources for the mental health professional, culture and children's intelligence* (pp. 255–263). Academic Press.
- van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 369–386). Routledge.
- van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, 13(1), 29–37.
- Vera, H. (2015). The social construction of units of measurement: Institutionalization, legitimation, and maintenance in metrology. In L. Huber & O. Schlaudt (Eds.),

- Standardization in measurement: Philosophical, historical, and sociological issues* (pp. 173–187). Routledge.
- Verschelden, C. (2017). *Bandwidth recovery: Helping students reclaim cognitive resources lost to poverty, racism, and social marginalization*. Stylus Publishing.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Walker, W. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (ETS Research Report No. RR-23-11). <https://doi.org/10.1002/ets2.12374>
- Warren, B., Vossoughi, S., Rosebery, A. S., Bang, M., & Taylor, E. V. (2020). Multiple ways of knowing: Re-imagining disciplinary learning. In N. S. Nasir, C. D. Lee, R. Pea, & M. M. de Royston (Eds.), *Handbook of the cultural foundations of learning* (pp. 277–294). Routledge.
- Webb, M. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test–III. *Educational and Psychological Measurement*, 68(2), 335–351.
- Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Lawrence Erlbaum Associates.
- Wertsch, J. V. (Ed.). (1985). *Culture, communication and cognition: Vygotskian perspectives*. Cambridge University Press.
- Wigfield, A., & Eccles, J. S. (1990). Test anxiety in the school setting. In M. Lewis & K. D. Rudolph (Eds.), *Handbook of developmental psychopathology* (pp. 237–250). Springer.
- Wilson, M., & Toyama, Y. (2018). Formative and summative assessments in science and literacy integrated curricula: A suggested alternative approach. In A. L. Bailey, C. A. Maher, & L. C. Wilkinson (Eds.), *Language, literacy and learning in the STEM disciplines: How language counts for English learners* (pp. 231–260). Routledge.
- Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin*, 76(2), 92–104.
- Wine, J. D. (1980). Cognitive-attentional theory of test anxiety. In I. G. Sarason (Ed.), *Test anxiety: Theory, research, and applications* (pp. 349–385). Lawrence Erlbaum Associates.
- Winstead, L. (2011). The impact of NCLB and accountability on social studies: Teacher experiences and perceptions about teaching social studies. *The Social Studies*, 102(5), 221–227. <https://doi.org/10.1080/00377996.2011.571567>
- Winter, P. C., Kopriva, R. J., Chen, S.-C., & Emick, J. E. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, 16(4), 267–276.
- Wolf, M. K., Kao, J., Herman, J. L., Bachman, L. F., Bailey, A. L., Bachman, P. L., Farnsworth, T., & Chang, S. M. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses: Literature review*. National Center for Research on Evaluation.
- Yaneva, V., & von Davier, M. (Eds.) (2023). *Advancing natural language processing in educational assessment*. Routledge.

- Young, J. W., King, T. C., Cogan Hauck, M., Ginsburgh, M., Kotloff, L., Cabrera, J., & Cavalie, C. (2014). *Improving content assessment for English language learners: Studies of the linguistic modification of test items* (ETS Research Report No. RR-14-23). ETS.
- Yu, A., & Ebbs, D. (2012). *Translation and translation verification*. TIMSS & PIRLS International Study Center.
- Zenisky, A., Hambleton, R., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, 9(1–2), 61–78. <https://doi.org/10.1080/10627197.2004.9652959>
- Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). *Using DIF dissection method to assess effects of item deletion* (College Board Report No. 2005-10). ETS.
- Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.

Part II

Foundational Concepts

Validity and Validation

Suzanne Lane

University of Pittsburgh (Emerita)

Scott F. Marion

National Center for the Improvement of Educational Assessment

Now, as a nation, we don't promise equal outcomes, but we were founded on the idea everybody should have an equal opportunity to succeed. No matter who you are, what you look like, where you come from, you can make it. . . . Where you start should not determine where you end up. (Obama, 2014, p. 1)

Validity frames all aspects of testing, from conceptualizing the targeted construct; to designing, developing, and implementing the test; to evaluating the consequences of test use. Fundamental to the validity of score meaning and test use are investigations of validity for different groups in the population defined by ethnicity, race, culture, language dominance, disability, and other relevant experiential, educational, social, and personal characteristics. Throughout the chapter, we typically refer to these various population groups as diverse groups. President Barack Obama's sentiment was a mandate for equal opportunity in higher education, but his words also are applicable to equal opportunity in testing, from accessing the construct to reaping any intended positive benefits of test use while minimizing any undesirable outcomes. Intentional test design that accounts for different ways of knowing and doing across diverse groups has the potential to improve the validity of score meaning and use for individuals—no matter who they are, what they look like, or where they come from.

We refer to *validity* as the extent to which theory and evidence supports or refutes proposed and actual or enacted test score interpretations, uses, and consequences and to *validation* as the process of gathering, evaluating, and synthesizing such evidence. Evaluations of validity depend on the specific uses of tests. For example, validity evidence to support or refute the use of a college-readiness test as a high school state assessment should be commensurate with this proposed purpose and not rely solely on evidence supporting the test for admissions decisions.

We adopt the argument-based approach to validity espoused by Cronbach (1988), Kane (1992, 2006, 2013), and Shepard (1993), which entails an interpretation and use argument and a validity argument, where the purpose, use, and context of testing are at the forefront of the argument-based approach. Emphasizing the need to focus on the validity of test use as well as score interpretation, Kane (2013) shifted from using the term *interpretive argument* to *interpretation/use argument*, although he emphasized use in his original framing of an interpretive argument. The interpretation and use argument (IUA) explicitly links the inferences from test performance to decisions and uses, including actions resulting from decisions. Consequences of test-based decisions and uses are therefore central to the argument-based approach to validity. The validity argument provides a comprehensive framework, including the rationales, assumptions, and backing or evidence, to support (or refute) the inferences and uses laid out in the IUA. Judgments regarding the evidence do not result in a yes/no endorsement regarding the validity of the score inference and test use; rather, they are more nuanced and suggest a degree of validity.

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014) referred to validity as the “degree to which evidence and theory support the interpretation of test scores for proposed uses of tests”

(p. 11) and to validation as a “process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed uses” (p. 11). We maintain that the *Standards* offers a narrow view of validity by indicating that validity refers to score interpretations “for proposed uses” instead of “and uses.” In contrast, we argue that validity refers to score decisions and test uses *including their consequences* and is not limited to the evaluation of score *interpretations* for a particular use. As an example, some tests are intended to serve as levers of educational reforms—such as raising academic standards and achievement for students, in particular, for historically underserved students—by signaling the importance of instructing students on rigorous content. The use argument for these tests as educational change agents requires a corresponding validity argument. Consequences, therefore, are central to the validity argument. Validation entails providing evidence regarding the use of tests and their intended and unintended consequences as well as evidence to support test score interpretations. The *Standards* are ambiguous regarding the use of tests and their consequences as part of the validity argument, whereas we support, and discuss in the chapter, the now long-standing position that test use and its consequences are integral to validity (Kane, 2006, 2013; Messick, 1989; Shepard, 1993).

We argue that fairness issues are validity issues because they are inherent in the validity of score meaning and test use for individuals from diverse groups. Consequently, because fairness issues are validity issues, they must be addressed in the IUA and validity argument. Fairness is at the core of all aspects of testing, from conceptualizing the targeted construct; to designing, developing, and administering the test; to evaluating the consequences of test use. Throughout this chapter, we use the argument-based approach to validity because it provides a mechanism to address validity and fairness throughout the testing process, from conceptualizing the construct to evaluating the consequences of test use.

As the title of this chapter implies, we discuss validity theory. However, we emphasize *validation*: the application of validity theory in conducting validity evaluations. We begin this chapter with the position that validity is an overarching concept that subsumes reliability/generalizability and fairness. We then review the evolution of validity as conceived by authors of the validity chapters in previous editions of *Educational Measurement* and by other validity theorists and researchers, providing a historical account of how consequences became central to validity and validation. We discuss the argument-based approach to validity espoused by Kane (2006) in the fourth edition of this text, as well as by other validity theorists and researchers, advocating its essential role for evaluating score meaning, test use, and resulting consequences. We then argue that consequences are essential to validity and the validation process and, further, discuss their importance in the IUAs and validity arguments for different testing programs, including credentialing testing programs, K–12 assessment and accountability programs, and language testing programs.

The next section addresses validity challenges when assessing individuals from diverse cultural and language groups. Our discussion supporting culturally responsive

and situated assessment practices that aim to provide more meaningful and equitable score interpretations and test uses for individuals from diverse cultural backgrounds is situated during a pivotal time of heightened recognition of the long-standing systemic inequities in our assessment and education systems. Our discussion of assessment and learning in context emphasizes the role of situated learning theory to improve the validity of test score inferences and uses. Further shifting the testing process to provide more valid score meaning and use for individuals from different cultural, social, experiential, and personal backgrounds requires partnerships among all those involved in and affected by testing. However, we recognize the challenges in designing such assessments and the need to obtain evidence to support their use.

We conclude by addressing applications of validity theory to testing programs, although throughout the chapter we emphasize applications of validity theory, including to K–12 assessment and accountability programs and to testing programs that use innovative technology. One of our main goals in this chapter is to wrestle with the tension between (a) the validity of score inferences and uses for assessments that are relevant and appropriate for all groups of individuals and (b) the large-scale assessment uses that rely on score meaning and test use that are essentially the same across groups.

VALIDITY AS AN OVERARCHING CONCEPT

Validity, reliability, comparability, and fairness are inherently interwoven, with validity as the overarching concept. In their discussion on the differing perspectives of fairness, Nisbet and Shaw (2019) concluded that validity, fairness, reliability, and comparability are “closely related and intersect at many points in most theoretical accounts” (p. 624). Moreover, as Messick (1994, p. 13) noted, “validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made.” Crisp (2017) reviewed the stance of validity theorists on the relationship between comparability and validity, advancing the position that comparability issues are part of validity as they relate to the appropriateness of score interpretation and uses of test outcomes. Historically, validity and fairness of score inferences and test use require the assumption of comparability.

Comparability and Validity

The validity of large-scale assessment score interpretations and uses is influenced by the comparability of score meaning for individuals as well as for aggregated groups, such as schools, states, cultural groups, English learners, and test takers with disabilities (Berman et al., 2020; DePascale & Gong, 2020; Keng & Marion, 2020). Comparability means that test takers who have the same test score have the same level of competence with respect to the construct a test is intended to measure (AERA et al., 2014). Investigations of score comparability have traditionally focused on statistical methods

to evaluate measurement invariance, such as multigroup factor analyses and differential item functioning (DIF).

Score comparability requires the critical assumption that the construct, as defined, is appropriate for all groups in the targeted population. If this assumption is untenable, the validity and fairness of score interpretations and uses will be compromised. As we will discuss later in this chapter, to attain a more responsive assessment of contextually bounded ways of knowing and doing, comparability investigations must begin with an appraisal of the appropriateness of the conceptualization and definition of the construct for all groups in the targeted population. Teachers, and others in classroom settings, may have to contextualize tasks to accommodate different backgrounds, cultures, and dominant languages to ensure valid interpretations and uses of assessment data. Since comparability, like validity and reliability, is evaluated in the context of use, comparability requirements for classroom assessments are not as stringent as those for educational large-scale monitoring or accountability tests, as well as other large-scale tests such as licensure and certification tests. In fact, classroom assessments (and, many would argue, other educational assessments) may better serve learning if they are tailored to individuals' social and cultural histories.

Fairness as Integral to Validity

The 2014 *Standards* (AERA et al., 2014) elevated fairness to one of the three foundational chapters, along with validity and reliability, for evaluating the quality and uses of tests and test score interpretations. The *Standards* posited that fairness is a “fundamental validity issue and requires attention throughout all stages of test development and use” (AERA et al., 2014, p. 49). As with generalizability and reliability, we argue that fairness forms the core of validity arguments, although we support the need to assert fairness's own identity so that equity and social justice issues in testing are brought to the forefront. Therefore, we will use both terms—validity and fairness—throughout this chapter.

Fairness issues are fundamental to the validity of test score interpretations and test use, and they surface throughout the testing process. Theory and logical and empirical analyses provide evidence to support or refute the validity and fairness of score meaning and test use for individuals with different social, cultural, and personal histories. Messick (1980) discussed the complexities in conceptualizing and labeling a construct based on theory and evidence: “[The] choice of construct labels is influenced by theory as well as by evidence and . . . by ideologies about the nature of humanity and society which add value implications that go beyond evidential validity” (p. 1019). Others have argued that it is not necessarily the targeted construct that is being assessed, but the individual's cognitive construction of knowledge and ways of knowing that arise from one's social and cultural experiences (Hood, 1998; Mislevy, 2018; Nortvedt et al., 2020).

Fairness evaluations typically address the validity of score inferences and test use for relevant groups in the targeted population; and because score inference and test use

are fundamental to validity, fairness, too, is fundamental to validity. As expressed by L. Shepard (personal communication, May 22, 2021), “first we (historically as a field) had validity, which was implicitly validity for the dominant group. Then, in response to the civil rights era, the field addressed validity for non-dominant groups and called that new thing ‘fairness.’” The groups identified for fairness evaluations arose from equity and social justice concerns. Whether based on gender identity, culture, ethnicity, language status, or disability, group definitions must attend to the heterogeneity of the individuals within each group. For example, English learners comprise many subpopulations, with various dominant languages as well as variants within dominant languages.

From a broader perspective, Rawls (2001) saw justice as fairness “to provide an acceptable philosophical and moral basis for democratic institutions and thus to address the question of how the claims of liberty and equality are to be understood” (p. 5). He advocated that every individual is entitled to the same basic liberties and equality of opportunities, and additional resources should be provided to those who do not have equal opportunities. Fairness in testing requires all test takers to have equitable opportunity to access, and demonstrate knowledge of, the construct measured so that score meaning and test use for individuals are comparable regardless of group affiliation. Messick (1998) argued that “evidence of fairness indicates that score meaning does not differ consequentially across individuals, groups, or settings; the major concern is comparability” (p. 1). We contend that evidence of fairness also indicates comparability and equity in terms of test uses and their consequences. Cronbach (1976) promoted the importance of values in the evaluation of test use, but he also acknowledged the importance of values in shaping validity evaluations. Messick (1998) expanded on this perspective, advocating that values, beliefs, and perceptions bear on fairness evaluations of testing programs: “Fairness judgments and validity judgments *are* value judgments” (Messick, 1998, p. 26).

Camilli (2006) provided a broad perspective of fairness, advocating that an evaluation of fairness may involve “legal, ethical, philosophical, and economic reasoning” (p. 225). He argued that concerns about fairness arise from the intended and unintended consequences, primarily adverse consequences of test use, and thereby noted that “investigations of fairness are framed by test use” (Camilli, 2006, p. 251). This view agrees with Messick’s (1998) assertion that fairness is an “overarching issue in the validity of test use” (p. 12). Camilli (2006) argued that issues of fairness are “inevitably shaped by the particular social context in which they are embedded” (p. 223).

Kane (2010) embraced these notions of fairness in his discussion of the relationship between fairness and validity. Fairness and validity, he maintained, are “closely related ways of examining the same question: Are the proposed interpretations and uses of the test scores appropriate for a population over some range of contexts?” (p. 177). He posited that this question is addressed from slightly different perspectives and entails different emphases. A fairness evaluation entails examining, across diverse groups, whether score interpretations are equally plausible, decision rules are equally appropriate, and consequences are comparable (Kane, 2010). Fairness is central to the evaluation of consequences, in particular, when they lead to differential outcomes across

groups. We support this perspective on fairness. And since consequential evidence is critical to validation, as we explain later in this chapter, we contend that fairness is at the core of validity arguments.

Fairness evaluations require “responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses” (AERA et al., 2014, p. 50).¹ A fair test

reflects the same construct(s) for all test takers, and scores from it have the same meaning for all the individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. (p. 50)

The *Standards* promotes sound perspectives on fairness and acknowledges fairness as fundamental to validity, but it falls short of explicitly addressing how fairness contributes to the validity of score decisions, test use, and the resulting consequences.

To summarize, we conceptualize fairness as a component of the validity of score interpretations and uses and the resulting consequences of test use. Fairness issues arise throughout the testing process, from the conceptualization of the construct to the consequences of test-based decisions. Fairness evaluations are critical for testing programs claiming to have indirect benefit beyond test score interpretations and score uses, such as tests used as levers to improve educational opportunities for students, especially students who have been historically underserved. Consequences can have differential effects, so this entails evaluating potential differential effects on relevant groups in the targeted population. Such differential effects may signal the systemic inequities of educational opportunities for certain groups of students. Validity rests on the fairness of consequences across relevant groups. Ethical and social justice considerations in the use of tests, and in evaluating their consequences, are integral to the validity argument.

Reliability/Generalizability as Integral to Validity

Reliability typically refers to the consistency, or invariance, of scores across replications of a testing procedure; it is essential to the validity argument. Consistency of scores may vary across samples of tasks, occasions, raters, contexts, and other relevant facets, and measurement error results from this inconsistency. Generalizability theory provides a useful framework for evaluating the generalizability of score inferences to the broader domain, and it allows for the estimation of error across relevant facets (Cronbach et al., 1972). Of interest is the generalization from a person’s test score to the person’s expected score over the hypothetical universe of tasks, raters, occasions, and other defined facets in the universe. Under classical reliability theory, scores are simply partitioned into true and error components. With generalizability theory, in contrast, analysis of variance techniques are marshaled to estimate separately the variability in a set of scores attributable to multiple sources of error, such as tasks, contexts, and occasions (Brennan, 1992; Cronbach et al., 1972).

The IUA typically makes claims about generalizing from the test score to the universe score (the score obtained over different samples of items, raters, occasions, and

other defined facets), and the validity argument calls for evidence to support the generalization inference. Generalizability of score inferences therefore is a necessary component of the IUA and the validity argument. In discussing reliability/precision of scores, the *Standards* state that the “level of reliability has implications for validity” and that reliability “of data ultimately bears on the generalizability or dependability of the scores and/or the consistency of classifications of individuals derived from the scores” (AERA et al., 2014, p. 34).

Cronbach et al. (1972) advocated a unified approach to validity and reliability, acknowledging the blurring of the distinction between them. In their text on the generalizability and dependability of measurements, they argued that the “theory of ‘reliability’ and the theory of ‘validity’ coalesce” (Cronbach et al., 1972, p. 234). As Kane (1982) suggested, “reliability is a necessary condition for validity because generalization is a key inference in interpretive arguments, but it is not a sufficient condition because generalization is not the only inference in the argument” (p. 529). Messick (1989) elevated construct validity to a superordinate role for evaluating score meaning and properties, including reliability. He indicated that an appropriate level of reliability depends on construct meaning; therefore, evidence of reliability consistent with the meaning of the construct is also evidence of construct validity (Messick, 1998). Following Cronbach and others, we contend that reliability and generalizability rightfully belong in, and are subsumed by, the validity argument.

EVOLUTION OF VALIDITY

This section traces the evolution of validity theory, emphasizing the seminal work of Cronbach and Meehl (1955) on construct validation and how it has shaped validity theory over the years. We also discuss the evolution of validity in the past editions of *Educational Measurement* as well as the *Standards*.

Validity for Test Use and Score Inferences

It has been long established that validity is a property not of the test, but of the test score inferences and test uses, with the acknowledgment that there could be multiple uses for a particular test. In 1946, Rulon asserted that it is unsatisfactory to view validity as the extent to which a test measures what it is purported to measure; rather, validity is concerned with “whether the test does the work it is employed to do” (p. 290) and it “depends on the propriety of the use made of the test” (p. 290). Rulon acknowledged that the type of validity evidence required depends on the particular use of a test: “We should not ask always for the same kind of proof of validity” (p. 291). In the first edition of *Educational Measurement*, Cureton (1951) reinforced the idea that validity refers to the purpose(s) of the test: “The essential question of test validity is how well a test does the job it was employed to do” (p. 621). The stance that validity is not an all-or-none proposition was explicitly stated by Cureton (1951): “The same test may be used for several purposes, and its validity may be high for one, moderate for another, and low for a third” (p. 621).

In their seminal article, Cronbach and Meehl (1955) advocated that validity refers to test score inferences and not the test itself: “One does not validate a test, but only a principle for making inferences” (p. 297). Referring to the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (American Psychological Association [APA], 1954), the first precursor to the *Standards*, Cronbach and Meehl (1955) indicated that validity is not an all-or-none proposition but, rather, that there are degrees of validity of score inferences: “If a test yields many different types of inferences, some of them can be valid and others invalid” (p. 284). They argued that the objective of validity inquiry is “to specify how one is to defend a proposed interpretation of a test” (p. 284). In the second edition of *Educational Measurement*, Cronbach (1971) maintained that a test is not validated, “but an *interpretation of data arising from a specified procedure . . . evidence that justifies one application [of test use] may have little relevance to the next*” (p. 447). He also saw validity falling on a continuum, in that “every interpretation has its own degree of validity; one can never reach the simple conclusion that a particular test ‘is valid’” (p. 447).

Criterion, Content, and Factorial Validity

The history of validity theory can be traced from the criterion model, to the content model, to, eventually, the construct model, with arguments for the sufficiency of one type of validity model for a particular test purpose to arguments for the need for multiple forms of validity evidence. In the early 1900s, validity was widely regarded as the correlation between test scores and a criterion measure—“validity coefficients”—which was considered sufficient validity evidence. Sireci (1998) cited Kelley (1927) and Thurstone (1932) as supporting correlational evidence as validity, but he also discussed their concerns, such as the relevance of the criterion to test use, the restrictiveness of an empirical approach, and the need to supplement validity coefficients with professional judgment.

Guilford (1946, p. 429) proposed two kinds of validity, where factorial validity involves an investigation of the internal structure of the test and practical validity calls for the correlation between a test score and a practical criterion. In discussing the limitations of professional judgment in test development and validation, Guilford argued that factor analysis can shed light on what is being measured. “In test construction and in job analysis,” he said, “things are not always what they seem” (p. 437). Thurstone (1952) extended his earlier thinking and, given the coarseness of criterion validity coefficients, argued for investigations of internal consistency when measuring cognitive functioning. Considering the psychological tests developed in the late 1880s and early 1900s, Gulliksen (1950) raised concerns about the assumption that a criterion has “intrinsic value.” He maintained that the appraisal of tests based on the relationship of test results with the expert judgment of a teacher or supervisor is fallible. He was concerned that this fallible judgment serves as a criterion, which is assumed to have “intrinsic value,” and that the test should approximate this criterion. Rather, empirical relationships between test scores and criterion measures should make theoretical sense in terms of what the test is intended to measure and what the criterion is presumed to encompass; Gulliksen warned, “[It is] dangerous to depend entirely on correlational validity to evaluate a test” (p. 514).

Regarding achievement testing, Rulon (1946) recommended that the appropriateness of the test be evaluated. He argued that the relationship between test scores and a criterion measure is not sufficient and that evidence to support validity depends on the purpose of the test. Rulon therefore proposed different types of evidence for different uses of achievement tests. As he discussed, when using a shorter version of an achievement test, the correlation between scores on an achievement test and a sampling of the items from the original test provides evidence for the shorter test as a replacement for the original test. For addressing whether the achievement test is appropriate for measuring student achievement, he proposed an evaluation of the content of the test, including the processes engaged in by the student. The question that needs to be addressed for this purpose of the achievement test is, “Are the materials of this test, and the processes called for on these materials, the same things and processes we are trying to teach the children?” (Rulon, 1946, p. 4). Support for this inquiry provides “the final proof of validity” (Rulon, 1946, p. 4). In his examples, Rulon proposed only one form of evidence to support a particular use of an achievement test.

Gulliksen (1950) argued for more than one form of evidence to support a particular test use, advocating, for example, that an evaluation of test content alone is not sufficient validity evidence. He suggested the use of factor analysis to better understand what is assessed, even if the results are inconsistent with expert judgment (because of the limitations of expert judgment). Loevinger (1957) supported the claim that considerations of test content alone are insufficient to establish validity. However, a decade later, Ebel (1961) dismissed the need for multiple forms of evidence for an achievement test, indicating that content validity is sufficient.

In the validity chapter in the first edition of *Educational Measurement*, Cureton (1951) initially defined validity as the “correlation between the actual test scores and the ‘true’ criterion scores” (p. 623), where the true criterion scores are a result of correcting the observed correlation for attenuation due to unreliability. He proposed two aspects of validity—relevance and reliability—where relevance is the “closeness of agreement between what the test measures and the function that is used to measure” (p. 622) and reliability refers to the consistency of scores. Later in his chapter, Cureton broadened his definition of validity by including logical relevance for a job knowledge test (i.e., job analysis), factorial relevance, and curricular relevance (or content validity) for an achievement test.

Emergence of Construct Validity

The *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954) proposed four types of validity: content, predictive, concurrent, and construct. Each type of validity required different types of evidence for different types of interpretations. The *Technical Recommendations* first proposed that construct validity be used only when “the test has no definitive criterion measure of the quality with which he is concerned and must use indirect measures to validate the theory” (APA, 1954, p. 54). However, the *Technical Recommendations* acknowledged that there may be more than

one purpose of a particular test and that the test's purpose sheds light on the type of validation needed. A year later, the *Technical Recommendations for Achievement Tests* (AERA & National Council on Measurements Used in Education, 1955) maintained that content validity is established by demonstrating that a test's items are a sample of the intended universe of items and advocated that establishing content validity was not sufficient for achievement testing. Rather, construct validity should play a prominent role in evaluating achievement tests. Predictive validity and concurrent validity are "criterion-oriented validation procedures" (Cronbach & Meehl, 1955, p. 282), where the criterion is obtained after the test is given for the former and, for the latter, the criterion score is obtained at the same time as the test score.

As indicated by Cronbach, chair of the initial APA Committee on Psychological Tests, and Meehl, a member of the committee, in their seminal article on construct validity, the "chief innovation" (Cronbach & Meehl, 1955, p. 281) of the 1954 *Technical Recommendations* was the inclusion of construct validity. The 1954 *Technical Recommendations* proposed that construct validity requires both logical and empirical investigation for validating the theory underlying the test by accumulating data to confirm predictions about the "variation of scores from person to person or occasion to occasion" (APA, 1954, p. 14).

Cronbach and Meehl (1955) extended the treatment of construct validity set forth in the 1954 *Technical Recommendations*.² Cronbach and Meehl defined a construct as a postulated attribute of individuals, and the goal of validation is to verify what constructs account for variance in test performance. They proposed construct validity for situations where content and criterion-related validity were inapplicable or insufficient for defining what the test is intended to measure. While establishing the importance of construct validity for personality tests, they also called for investigations of which constructs account for test performance on other tests (e.g., measures of aptitude, achievement, and interests) in addition to other forms of validity evidence, because different types of validity evidence may provide meaning for scores for a particular test.

Construct validity, along with criterion-related (predictive and concurrent) and content validity, was adopted by the 1966 *Standards for Educational and Psychological Tests and Manuals* (AERA et al., 1966). The 1966 *Standards* replaced the 1954 and 1955 *Technical Recommendations*.

Construct Validation as Scientific Reasoning

Construct validation, as scientific reasoning, begins with an investigator's claim that the test measures a construct that has certain meanings, leading to identifying and testing hypotheses to confirm or disconfirm the claim (Cronbach & Meehl, 1955, p. 290). This position is consistent with the view of Campbell (1957) that researchers are obliged to consider rival hypotheses, rather than taking a confirmationist perspective, when examining the validity of causal inferences from the results of experiments. The validation of

the claim that the test measures a construct requires a nomological network, “a system of laws [propositions] which constitute a theory” (p. 290), that represents the construct meaning. “[The] laws may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical concepts to one another” (Cronbach & Meehl, 1955, p. 290). To provide construct meaning, the propositions relate the test to the construct, the construct to other constructs, and the construct to observable variables. As summarized by Cronbach and Meehl (1955), a “construct is defined implicitly by a network of associations or propositions in which it occurs” (pp. 299–300), requiring a strong chain of inference to demonstrate that a test, in fact, measures the construct. One uses inductive argumentation to verify the network by demonstrating that the construct measured is consistent with the evidence (Cronbach, 1971), and alternative interpretations are considered and evaluated (Cronbach & Meehl, 1955). Cronbach and others underscored the need to include criterion measures in the construct validation framework to ensure that the evaluation of the criterion measures are held to the same psychometric standards as other measures (Cronbach, 1971; Cronbach & Meehl, 1955; Messick, 1965). Although they were suggested by Cronbach and Meehl (1955), Campbell and Fiske (1959) formalized multitrait, multimethod approaches for simultaneously evaluating convergent and discriminant evidence in construct validation.

Cronbach and Meehl (1955) argued for the importance of documenting construct validation research, including delineating the proposed interpretation, how well the interpretation is substantiated, and the evidence and reasoning to support the claim, as well as including alternative interpretations that were considered and rejected. They emphasized the need to specify what parts of the proposed interpretations are “speculations, extrapolations or conclusions from insufficient data” (Cronbach & Meehl, 1955, p. 297). Both logical and empirical evidence have a role in construct validation. The 1954 *Technical Recommendations* outlined different types of construct validation studies, including examining factor structure, whether experimental manipulation induces changes aligned with theory, and whether test scores are related to similar and dissimilar measures. Cronbach and Meehl (1955) broadened the scope to include studies of change over occasion (stability measures) as well as studies of processes engaged by the test taker.

Loevinger (1957) proposed that construct validity has three components—substantive, structural, and external—which are mutually exclusive, exhaustive, and mandatory. The substantive component of validity encompasses content validity and refers to the “extent to which the content of the items included in (and excluded from) the test can be accounted for in terms of the trait believed to be measured and the context of measurement” (Loevinger, 1957, p. 661). The structural component refers to the “extent to which structural relations between test items parallel the structural relations of the manifestations of the trait being measured” (p. 661). The external component refers to the relationship between the test score and other variables (including relevant criteria) with assumed relationships (including no relationship) with the test

score. The content of the items should be consistent with the proposed interpretation, the structural relations of the items should be consistent with the structural relations of other measures of the same trait, and the external correlations of the test score should be consistent with predictions based on knowledge of the proposed trait (Loevinger, 1957).

The validity chapter by Cronbach (1971) in the second edition of *Educational Measurement* elaborated on Cronbach and Meehl (1955) regarding validation as scientific inquiry into score meaning. The 1971 chapter reinforced the proposition that validation is an inductive argument used in the verification of the proposed network, which “is a representation of present knowledge, not the only possible representation” (Cronbach, 1971, p. 483)—hence, the need for proposing and testing counterhypotheses. Construct validation begins with a statement of the proposed interpretation, which will suggest the evidence needed to support the interpretation. It is a process of “reasoning from data to conclusions” (Cronbach, 1971, p. 484). Messick (1975) advocated for this conceptualization of construct validation, which “links a particular measure to a more general theoretical construct, usually an attribute or process or trait, that itself may be embedded in a more comprehensive theoretical network” (p. 955). Both the 1966 *Standards* and the 1974 edition (AERA et al., 1974) discussed validity in terms of criterion-related, content, and construct validity. The 1974 *Standards* proposed that validity questions are of two types: those dealing with the intrinsic nature of the meaning of the measure and those dealing with the test score as an indicator of other variables. Messick (1980) endorsed this perspective: “Evidence bearing on the meaning of the measure embraces all of construct validity, whereas evidence for certain predictive relationships contributes to that part called nomological validity” (p. 1016). Similarly, Embretson (1983) distinguished between nomothetic span and construct representation, where the former is an empirical network of relationships of the test with measures of other constructs and criterion measures, while construct representation involves the identification of theoretical components that underlie task performance, including processes, strategies, and knowledge. Partitioning the task into these features allows for the building of these components into a process model that provides meaning to the construct (Embretson, 1983).

Cronbach (1989) later reflected on his and others’ confirmationist perspective on construct validity, with its focus on accumulating evidence in support of the interpretation rather than evidence in support of plausible rival hypotheses. He encouraged the use of both empirical and logical methods to test plausible rival hypotheses as part of validation. Messick (1980) also promoted the perspective that both confirmatory and disconfirming strategies should shape construct validation studies. He advocated that “corroborative evidence provisionally supports both the construct theory and the test interpretation, while disconfirming evidence undercuts either the theory or the measure, or both” (Messick, 1980, p. 48). Cronbach (1989) further advocated that validation is a “community process”; external critics with differing perspectives need to play a role in the validation process. We agree, and we later argue that multiple stakeholder groups

should have a voice in shaping both the IUA and the validity argument, in part, to address equity and social justice issues.

Construct Validity as a Unitary Concept With Types of Evidence

Construct validity as a unitary concept surfaced in the work of Loevinger (1957), who urged the measurement community to view “psychometrics with construct validity as its central concept” (p. 689). She argued that construct validity is the “whole of validity from a scientific point of view” (p. 636), but acknowledged other forms of evidence, such as predictive validity coefficients, for more practical applications. Cronbach and Meehl (1955) also spoke to a unified validity, noting that many forms of evidence are relevant to construct validity, including “content validity, inter-item correlations, inter-test correlations, test-criterion correlations, studies of stability over time, and stability under experimental intervention” (p. 300). In a similar vein, the 1954 *Technical Recommendations* stipulated that content validity and criterion-related validity provide evidence in support of construct validity.

Messick (1980) solidified the unitary view, advocating that validity is “an overall degree of justification for test interpretation and use” (p. 1014). He further argued that “different kinds of inferences from test scores require different kinds of evidence, not different kinds of validity” (Messick, 1980, p. 1014), thereby moving the discussion from types of validity to types of validity evidence. The *Standards* endorsed this unified view with the 1985 edition, referring to validity as the “appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (AERA et al., 1985, p. 9), where “there is often no sharp distinction between test content and test construct” (p. 11). There are not different types of validity; rather, there are different types of validity evidence: construct related, content related, and criterion related (AERA et al., 1985).

Messick (1989) argued for the need to integrate and synthesize different forms of validity evidence because they “are interrelated operationally and logically” (p. 26). Both the 1999 and the 2014 *Standards* endorsed the integration of different forms of evidence: “A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses” (AERA et al., 1999, p. 17; 2014, p. 21).

Messick’s Unified Validity

In the third edition of *Educational Measurement*, Messick (1989) characterized validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13). Messick (1980, 1989) employed a progressive matrix as a heuristic to represent two sources of justification of testing crossed with two outcomes, or functions, of testing (Table 4.1). The two sources of justification of testing are the evidential basis of score meaning and the consequential

Table 4.1 Messick's Facets of Validity

Basis	Test Interpretation	Test Use
Evidential basis	Construct validity	Construct validity + score relevance + score utility
Consequential basis	Construct validity + value implications	Construct validity + score relevance + score utility + value implications + social consequences

Note. Adapted from "Validity," by S. Messick, 1989, in R. L. Linn (Ed.), *Educational Measurement* (3rd ed., p. 20), American Council on Education and Macmillan.

basis of score valuation. The two functions of testing are test interpretation and test use leading to validation processes, which include meaning and values as well as test interpretation and test use. Expanding on the view that construct validity is a unified validity, Messick (1989) argued that construct validity binds the validity of test interpretation to the validity of test use.

Messick extended Cronbach and Meehl's perspective on validity as a unitary concept. He proposed his unitary conceptualization of validity because he was concerned with what he considered weak validation practice driven by the tripartite approach to validity and validation. For example, he was concerned that evaluators were claiming that strong content-related evidence for achievement tests was sufficient for establishing the validity of score inferences, without considering the threats to validity arising from other sources, such as poorly written items, the ways in which test takers interact with the items, test-taker motivation, and distractions during testing.

Messick's conceptualization of construct validity is the core of the interpretation and use functions, underscoring score meaning and values in both test interpretation and test use. Construct validity is represented in each cell of Table 4.1, along with aspects that depend on the particular justification and use of testing. The evidential basis of test score interpretation is construct validity, providing evidence to support score meaning. The evidential basis of test use includes construct validity along with the relevance of scores for the given test purpose and the utility of the scores in the testing application. The consequential basis of test interpretation is the appraisal of value implications of score meaning, including value assumptions associated with the construct label, the theory underlying the construct and score meaning, and the ideologies that provide the theoretical perspective. The consequential basis of test use accumulates all aspects of the progressive matrix and is the appraisal of potential and actual social consequences of testing (Messick, 1989). Intended and unintended consequences of testing represent intended and unintended action implications of score meaning. Adverse consequences bear on the meaning of test scores and, therefore, on the meaning of the construct. Messick (1989) advocated that information about the utility of scores for a given testing purpose is both validity evidence and a value consequence.

Kane (2012) expressed concern with Messick's unitary approach to validity because it does not lend itself to a pragmatic approach for validity investigations. The unified

view is “conceptually rich and suggestive, but it is not easy to implement effectively, because it does not provide a place to start, guidance on how to proceed, or criteria for gauging progress and deciding when to stop” (Kane, 2012, p. 8). His development of the IUA and validity argument provides a more pragmatic approach to validation. Shepard (1993) agreed with Messick’s conceptualization of validity but disagreed with the segmented presentation of the facets, arguing that the “matrix was a mistake because it led to false dichotomies between aspects of validity that are inherently entwined” (p. 269). As an example, she argued that the matrix could lead to potential misinterpretations, such as thinking that values are only relevant for the consequential basis. In fact, values are inherent in the evidential basis of test interpretation, where constructs are conceptualized and labeled. What was intended as a helpful heuristic has led to considerable controversy in educational measurement. For example, Cizek (2012, 2020) argued for separating the interpretation and use columns, while Newton and Shaw (2016b) argued for separating questions of social value from those of technical quality. We agree with Shepard (2016, p. 275) that “Messick’s two columns were intended to deepen our thinking about each aspect, but these aspects or facets cannot be addressed independently.”

Values and Ethics in Testing

The measurement of psychological and educational constructs is value laden (Cronbach, 1976; Messick, 1975). Deciding whether “a test should be used for a particular purpose raises the issue of values in measurement” (Messick, 1975, p. 962) because test use is an “ethical issue” (Messick, 1980, p. 1020). In his early writings on personality tests, for example, Messick (1965) argued that whether to use a test for a particular purpose may require justification that draws on social values. Ethical issues arise in the use of personality measures for selection, and the choice of a measure involves an inherent value judgment; it involves decisions regarding what personality characteristics are desirable, which may result in reduced diversity. The field has become more aware of, and sensitive to, this issue, with the recognition that many personality traits such as assertiveness as well as other constructs are culturally shaped.

Messick (1989) later argued that validity and values are one. Validity is an “ethical imperative of testing” (Messick, 1980, p. 1020), involving judgments about accumulated evidence. Validity entails an evaluation of the value implications of test interpretation and use, and these value implications derive primarily from the test’s construct meaning; however, they also derive in part from ideologies of social science, education, and social justice, and therefore they go beyond construct meaning. Test use rests on assumptions about possible test consequences—both intended and unintended—of score-based decisions and on the values associated with these different consequences (Messick, 1975, 1980, 1989). Messick (1994) contended that validity, reliability, comparability, and fairness involve evaluative judgments that reflect social values.

Decisions made in test design, development, and use rest on value judgments (Kane, 2006; Messick, 1989). There are competing, and potentially conflicting, values that

bear on validity, fairness, and social justice issues. Value judgments are involved in the articulation of the test's intended purposes and uses, the conceptualization of the construct to be assessed, the design and development of items and scoring rules, the development of performance standards for determining competency, and the design of score reports and accompanying interpretive guides (Kane, 2006). Importantly, values play a significant role in translating the targeted construct (or knowledge skills, and other attributes) into test specifications because of the prioritization inherent in this test design exercise. Value judgments are inherent in the articulation of the intended positive consequences and in the search for unintended negative consequences. Consider the case of using tests in decisions regarding high school graduation. This use was intended to raise standards and improve learning, but it had a differential effect on schools and groups of students—in particular, underserved students. Judgments regarding whether the benefit of using tests in graduation decisions to raise standards outweigh their undesirable effect on different groups of students vary depending on the beliefs, values, and assumptions held by different stakeholders (Lane, 2020). Zwick (2018) warned that using college and career on-track labels, such as a “not ready” designation, may result in lower educational expectations and, consequently, lower student outcomes, especially for underrepresented groups in higher education.

Threats to Validity

As described by Messick (1989), construct underrepresentation and construct-irrelevant variance are two potential threats to the validity of score interpretations and uses. Construct underrepresentation occurs when the assessment is too narrow and fails to capture important facets of the targeted construct. Construct-irrelevant variance occurs when the assessment scores reflect factors that are irrelevant to the targeted construct or when other extraneous factors (e.g., language barriers, test speededness) influence test scores. Construct-irrelevant variance negatively affects the scoring inference, which in turn threatens the validity of decisions and uses. Construct underrepresentation likely undermines generalizability claims and extrapolations from test performance to the real-world criterion, which also threatens the validity of decisions and uses. These two threats to validity play a central role in the evaluations of the validity of score meaning and use for all testing programs. We discuss these and other threats to validity and fairness throughout this chapter.

ARGUMENT-BASED APPROACH TO VALIDITY AND VALIDITY EVIDENCE: FROM THEORY TO IMPLEMENTATION

Previous chapters in *Educational Measurement* discussed the current conceptualization of validity and validation. A key shift between Messick (1989) and Kane (2006) was in Kane solidifying an argument-based approach to validation. While Messick's (1989) seminal chapter crystalized the unified view of validity, the theoretical and philosophical

orientation in his chapter left many wondering how to go about evaluating the validity of test score inferences and uses. Kane's (1992, 2006, 2013) argument-based approach to validation was a practical way to carry out validity evaluations. The terms *interpretation and use argument* and *validity argument* have been pervasive in the literature since 2006, and these concepts are well represented in current validity evaluations. This section reviews how validity arguments have been conceived, how this conceptualization has evolved (e.g., Kane, 2013), and how it has been applied. We discuss how the increasing use of principled approaches to test design and development intersects with argument-based approaches to validity. We conclude this section with an exploration of the use of theories of action in test validation and how such theories of action can bridge test validation and program evaluation.

Validation and Argumentation

The beginnings of an argument-based approach can be seen as far back as Cronbach and Meehl (1955): "We seek to specify how one is to defend a proposed interpretation of a test: *we are not recommending any one type of interpretation*" (p. 284). But Cronbach (1971) generally was seen as advocating a scientific inquiry approach to validation rather than having the evaluative focus proposed by Kane and others. Haertel (2018) characterized Cronbach (1971) as considering validation a scientific inquiry activity, but noted that Cronbach invited value judgments by, for example, asking developers to consider whether the names they gave their test allowed for overly broad score interpretations.

Following publication of Messick's (1989) validity chapter, researchers and practitioners began wrestling with how to conduct validity evaluations in practice that were faithful to his unified conception of construct validity. Shepard (1993), reaching back to Cronbach (1971, 1988), advocated a straightforward means to prioritize validity questions. Using an evaluation framework, she proposed that validity studies be organized in response to the questions, "What does the testing practice claim to do? What are the arguments for and against the intended aims of the test? and What does the test do in the system other than what it claims, for good or bad?" (Shepard, 1993, p. 429). These questions are directed at concerns about the construct relevance, interpretation, and social consequences, respectively. This framing is consistent with Kane's approach.

Kane's Framework

Kane (2006) proposed using two types of arguments:

An interpretive *argument* specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances, [whereas] the *validity argument* provides an evaluation of the interpretive argument. (p. 17)

Kane (2013, 2016) later renamed his interpretive argument the interpretation and use argument, signifying the inseparability of score interpretations and test uses. Kane (2016)

summarized this updated view of validation succinctly: “I think of validity as the extent to which the proposed interpretations and uses of test scores are justified” (p. 198).

The IUA outlines what the user/evaluator believes should occur—and why—as a result of the testing and related systemic data collection, whereas the validity argument involves weighing the available evidence and logic to decide whether the specific interpretations and uses are justified. Explicitly specifying the proposed interpretations and uses of the assessment, developing a measurement procedure consistent with these proposed uses, and then critically evaluating the plausibility of the initial assumptions and resulting inferences is more straightforward than evaluating the validity of an assessment under a construct model. Kane was not abandoning Messick’s unified view of construct validity; rather, he was providing a more pragmatic approach for evaluating the validity of score inferences and uses than under a strict construct model. The construct model is based on a research approach in which one searches for causal connections (e.g., Markus & Borsboom, 2013), whereas Kane’s argument-based approach works from an evaluation perspective to determine whether an assessment program is operating as planned, with minimal unintended negative consequences.

Kane (2006, 2013) pushed for the development of the IUA in the assessment design phase. Specifying purposes and uses up front and then designing an assessment to fit these intentions is embedded in principled approaches to test design (e.g., Mislevy & Haertel, 2006; Mislevy et al., 2003; see also Huff et al., this volume). Such design approaches require a detailed specification of the assessment argument early in the design process. This works well for new tests, but there has not been sufficient attention to validity in the design of many of the current state and certification assessments; consequently, evaluators likely have to retrofit a validity argument to the existing testing system. This is not an excuse to avoid validity evaluations, but a recognition that it is more challenging to do so after the fact.

The IUA

The IUA is essentially a mini-theory that provides a framework for making sense of score inferences and test use. Like a theory, the IUA guides data collection and methods for conducting validity analyses. And IUAs, like theories, also are falsifiable. Thus, validation is not a confirmationist exercise. It is helpful to think of the IUA as a series of if–then statements, such as, “If the student scores at the proficient level on the state achievement test, then the student has a solid understanding of the required grade-level content and skills.” Of course, there are many more if–then statements for any given test and test scores across and within the different inferences.

Kane (2006, 2013) noted two stages of the IUA. The development stage focuses on the development of measurement tools and procedures, as well as the corresponding IUA. Kane suggested that it is appropriate to have a confirmationist bias in this stage—a stance favoring evidence and interpretations that support the current state of the assessment program—because the developers are trying to make the program work as well as possible. Kane also argues that, in the development stage, likely challenges to the IUA

can be identified and addressed. The appraisal stage, in contrast, focuses on the critical evaluation of the IUA, where the most questionable assumptions and the intended and unintended consequences are deliberatively evaluated.

One of the most effective challenges to IUAs is to propose and substantiate alternative claims (or plausible rival hypotheses) that are equally or more plausible than the proposed proposition. Using the earlier example, an alternative inference could be as simple as, “If the student scores at the proficient level on the state achievement test, it indicates the student benefited from their background knowledge and out-of-school enrichment experiences.”

From a pragmatic level, test validation is the process of offering assertions (propositions) about a test or a testing program and then collecting data and posing logical and empirical arguments to challenge those assertions. Using the assertion and alternative hypothesis in the previous example, the evaluator should design studies that evaluate the degree to which test scores are sensitive to differences in the quality of instruction compared with students’ background knowledge and alternative hypotheses. In essence, validity evaluators are continually trying to challenge the claims being made.

Enacted Interpretations and Uses

The IUA focusses largely on *proposed* interpretations and uses (Kane, 2013). O’Leary et al. (2017) argued that *enacted* test score interpretations and uses must be part of the validity evaluation. Haertel (2018) suggested that following these IUAs to their logical conclusions must lead to enacted interpretations rather than intended score interpretations only. Haertel (2018) pointed out that misinterpretations associated with standards-based reporting practices (e.g., *Proficient*) are a key example supporting the focus on enacted interpretations and uses in addition to intended ones. The documentation and validity evidence associated with defining achievement levels is often lacking (e.g., National Research Council [NRC], 2017), particularly in terms of the enacted interpretations and appropriate uses of the achievement levels.

Focusing on enacted interpretations and uses clearly establishes the centrality of evaluating potential unintended consequences in validation efforts. Kane (2016), in his more recent writings, shifted to using the term IUA to stress the importance of test use and resulting consequences, although his previous conceptualization of the interpretive argument did embody use. Kane (2016) discussed the difference between Messick’s more limited and Cronbach’s more expansive view of the role consequences in validation, leading up to his discussion of the interpretation and use model:

The interpretation-and-use model adds an element to the consequences-as-indicators model. It adds Cronbach’s notion that negative consequences count against the validity of test use, even if they are not attributable to any flaw in the test or testing procedures, to Messick’s notion that negative consequences that are attributable to a flaw in the test or testing procedures count against the validity of the test score interpretation. (pp. 207–208)

Kane (2016) argued further that “the responsibility for evaluating specific claims would fall on those who make the claims, explicitly or implicitly. . . . To make consequential decisions without bothering to consider the consequences of these decisions would seem to be irresponsible” (p. 209).

Sources of Validity Evidence and the IUA

To frame validation efforts, Messick (1989) proposed six aspects of validity:

- content: evidence of content relevance, representativeness, and technical quality
- substantive: theoretical rationales for consistencies in test responses, including cognitive process models of task performance
- structural: evaluations of the appropriateness of the scores relative to the structure of the targeted construct(s)
- external: convergent and discriminant evidence from comparisons, as well as evidence of test–criterion relationships
- generalizability: evidence of the extent to which score properties and inferences generalize both to and across population groups (i.e., fairness), settings, contexts, and tasks; evidence of test–criterion relationships
- consequential: the value implications of score interpretation as a basis for action, as well as the actual and potential consequences of test use, particularly regarding sources of test invalidity related to issues of fairness and social justice

The 1999 and 2014 *Standards* adopted a similar frame for discussing validity evidence by proposing five types of evidence: evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing.³ The forms of validity evidence proposed by either the *Standards* or Messick (1989) provide a structure for evaluating test score interpretations and test uses. Researchers (e.g., Sireci, 2013) have suggested using these sources of validity evidence as a framework for validation research; however, we agree with others that these sources are insufficient (Chapelle, 2011; Newton, 2016). As Chapelle (2011) warned, validation involves examining the inferences in the argument instead of merely consulting a list of types of validity evidence. Similarly, Newton (2016) argued that the “five sources framework provide an impoverished view of validation evidence, which thereby risks practitioners designing suboptimal validation research programs” (p. 9).

The IUA provides a comprehensive framework for validation by linking intended inferences and uses to evidence, as well as linking plausible alternative inferences and uses to evidence. Kane (2006) offered examples of inferences that could be applied to many testing programs:

- scoring inferences
- generalization inferences
- extrapolation inferences
- decision/use inferences

These inferences involve, respectively, evaluating the link between test performance and a test score; generalizing the test score to the expected score over the hypothetical universe of items, testing occasions, raters, and other relevant facets; extrapolating from the hypothetical universe to the broader domain of practice; and making appropriate decisions and uses.

While often interpreted as a conjunctive set of inferences that must be evident in all testing programs' IUAs, these instead were meant to illustrate potential inferences to investigate (Kane, 2006, 2016). The IUA must outline the inferences and intended uses specific to each testing program.

In addition to Kane's four inferences, we propose a performance inference that links the tasks to the test taker's performance. This requires evidence to support the claim that tasks elicit the intended knowledge, skills, and other attributes and that the assessment is representative of the construct. This explicitly positions the conceptualization of the construct and test design at the forefront of the IUA and validity argument. In the context of language assessment, Chapelle et al. (2007, 2010) expanded Kane's IUA by adding the domain description prior to the scoring inference, which is similar to the inclusion of construct representation by Shaw et al. (2012). The evaluation of the construct representation of tasks in terms of theoretical attributes (e.g., processes, strategies, and knowledge) underlying item performance proposed by Embretson (1983), as well as principled approaches to test design, provide backing in support of the performance inference. As the first link in the interpretive chain, the performance inference is made from the task to performance, with the warrant for this inference being that the tasks elicit performance that represents the intended construct and the assessment is representative of the construct.

To be sure, the *Standards* outline critical sources of evidence for evaluating the validity of test score interpretations, but they provide little guidance for organizing, implementing, and synthesizing validity evaluations. Therefore, an IUA should be used to frame validity evaluations while integrating the *Standards*' five sources of evidence, as well as other sources of evidence as appropriate. We now consider in more detail the performance inference along with Kane's four inferences, providing examples of the types of evidence that could be used to support the various claims.

Performance Inferences

The performance inference explicitly positions assessment design and development within the IUA and validity argument. There are many tools and processes for evaluating the performance inference that map nicely onto the categories of the *Standards*. For example, evaluations of test content—the degree to which the items and tasks are representative of the targeted construct and whether the targeted construct is appropriately represented by the items and tasks—is critical to the performance inference. Evidence related to response processes has received little attention as being critical to the score inferences until recently, but such evidence now is regarded as fundamental to evaluating the performance inference (Ercikan & Pellegrino, 2017; Leighton, 2017).

The following statements are examples of the types of evidence that support performance inferences while drawing on sources of evidence described in the *Standards*.

- The assessment tasks evoke the intended knowledge, skills, strategies, and processes.
- The assessment is representative of the intended construct and does not assess irrelevant constructs.
- Test takers' cultural, linguistic, and disability characteristics do not limit access to, and demonstrate knowledge of, the targeted construct.
- Cognitive models used in test design are appropriate for different groups in the targeted population.

Score Inferences

According to Kane (2016),

The validation of a score interpretation involves an investigation of whether the scores mean what they are supposed to mean, and the interpretation is said to be valid if the claims inherent in the interpretation are supported by appropriate evidence. (p. 199)

Scoring inferences form the basis of almost any validity evaluation because if the scores do not reflect the intended evaluations of test-taker performance, test score inferences are unwarranted.

Kane's articulation of the scoring inference receives little attention in the *Standards*. For example, validity evaluators likely would state the following scoring claims, but they do not relate directly to any source of evidence outlined in the *Standards*.

- The scoring rules are appropriate for defining differences in score levels for the constructed-response items.
- The scoring rules are consistently and accurately applied to all test takers (i.e., there is no bias in the scoring rules).
- There are adequate procedures for ensuring consistency and accuracy of scoring.
- The procedures used to link scores across various test administrations are employed appropriately to ensure comparable score meaning across administrations.
- There is a clear description regarding the quality of the selection, training, and monitoring of scorers.

Generalization Inferences

Generalizability theory is a sampling theory, where the assessment is viewed as a sample from a universe of generalization (Brennan, 1992; Cronbach et al., 1972). Generalizability theory addresses the accuracy with which we can generalize from a person's score on a particular set of test items to the person's universe score over the population of items, raters, occasions, and other specified facets (Brennan, 1992; Cronbach et al.,

1972). There are extensive tools and processes for evaluating error (measurement and sampling) and reliability, both of which are subsumed by generalizability theory. Mapping to the *Standards*, generalizability requires a careful specification of the universe of generalization, which can be part of the evidence related to test content. The following statements are examples of the types of evidence an evaluator would collect to support inferences related to generalizability.

- There is evidence related to the test's internal structure, that is, how items fit together or how test-taker performance generalizes across items and/or domains (e.g., factor analysis results).
- There is evidence that the internal structure is similar across identifiable groups (e.g., cultural, racial, language) to support generalization across groups.
- There is evidence that accommodations yield valid score inferences.
- There is evidence of an appropriate approach for evaluating various errors associated with test scores (e.g., interrater consistency, person-by-item interactions).
- There is evidence that various sources of error associated with items and tasks are minimized.
- There is evidence that test scores have the appropriate level of reliability/generalizability for the intended purposes and uses of the scores.

Extrapolation Inferences

Many validity evaluations either ignore or downplay extrapolation—making inferences from the hypothetical universe to the broader domain of practice—in part because it is challenging to provide convincing evidence of it (e.g., Haertel, 2018). Nevertheless, evidence based on the relationships between test scores and other variables can help evaluate inferences regarding whether test scores represent the larger domain beyond the universe of generalization—the real-life criterion. Such evidence includes concurrent and predictive validity evidence, because both establish the relationship between test scores and a criterion representing the domain of interest. For college entrance examinations, for example, the predictive relationship with the criterion—first-year grade point average (GPA), typically—is a critical aspect of the validity argument for using such tests. For certification tests, content experts' review of the knowledge, skills, and other attributes (KSAs) tested in relation to the KSAs used in practice provide evidence for the extrapolation inference. The following statements are examples of the types of evidence that support extrapolation inferences.

- There is evidence of the expected relationships of test scores with results from tests hypothesized to measure both similar constructs and dissimilar constructs.
- There is evidence that the relationships between test scores and results from other measures are similar across identifiable groups (e.g., cultural, racial, language).
- There is evidence from expert judgment and other studies that the KSAs measured by the test are required for success in the targeted domain, such as for a particular job or subsequent education.

- There is evidence demonstrating the relationship between scores on college admissions tests and success in postsecondary education.
- There is evidence that test results reflect skills critical to successful job performance.

Decision/Use Inferences

Kane (2016) endorsed consequences as part of validity evaluation, in that evidence related to consequences is necessary for evaluating decisions and uses: “A score use involves a sequence of claims that culminates in a decision, and decisions are justified in terms of their expected consequences” (p. 198). As such, evidence related to consequences plays a critical role in evaluating the validity of decision/use inferences. Evaluators examining the plausibility of decision/use inferences likely would collect evidence along the lines of the following examples.

- There is evidence that supports the standard-setting process (e.g., documentation of process and materials used, surveys of panelists).
- There is evidence that the inferences associated with the labels (e.g., *Proficient*, *Basic*), descriptors, and interpretations of the achievement levels resulting from the standard-setting process are justified and understood.
- There is evidence that classification decisions based on test scores (e.g., *Proficient* or *Basic*, *Competent to Practice* or *Not Competent to Practice*) are supported through multiple types of studies (e.g., decision consistency and accuracy, content alignment, performance in the next grade).
- There is evidence that classification decisions based on test scores (e.g., *Proficient* or *Basic*, *Competent to Practice* or *Not Competent to Practice*) do not lead to unintended negative consequences, such as restricted or denied opportunities or negative labeling.
- There is evidence that decisions based on test scores (e.g., college admissions) are supported and do not lead to unintended negative consequences for identifiable groups of students.

Synthesizing Validity Arguments

Construct validation is . . . never complete. Construct validation is better seen as an ever-extending inquiry into the processes that produce a high or low test score and into the other effects of those processes (Cronbach, 1971, p. 452).

Cronbach’s proclamation may have discouraged users from engaging in validation for fear of getting stuck in never-ending studies. An advantage of Kane’s argument-based approach to validation is that it provides considerably more structure to validity evaluations than Cronbach’s construct model. Similarly, Haertel (1999) emphasized that individual pieces of evidence do not render an assessment system valid or not; rather, evidence and logic are synthesized to evaluate the IUA. Such activities should allow users and others to make overall judgments about the efficacy of the testing program. Yet, there are few examples, in practice or the literature, of validity argument syntheses.

The *Standards*, in a short section at the end of the validity chapter, provides general guidance for integrating the various sources of evidence into a summary judgment:

A test interpretation for a given use rests on evidence for a set of propositions making up the validity argument, and at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible. At some point the effort to provide sufficient validity evidence to support a given test interpretation for a specific use does end (at least provisionally, pending the emergence of a strong basis for questioning that judgment). (*Standards for Educational and Psychological Testing* (AERA, APA, NCME), 2014, p. 22)

Unfortunately, the *Standards* does not provide guidance for how evaluators should weigh the various sources of evidence when making a judgment about test use. For example, does response-process evidence count more than test score relationships with other variables (assuming both are relevant to a particular test use)? Kane's argument-based approach appears fairly conjunctive because the evaluator weighs evidence related to performance, scoring, generalization, extrapolation, and ultimately decision/use inferences—a high bar, to be sure.

Integrating and synthesizing various types of empirical evidence and logical arguments into an evaluative judgment can be a challenging activity. In operational assessment programs, furthermore, new and varied information comes in over an extended time interval (particularly consequence evidence), which exacerbates this challenge. Kane (2006) anticipated the timing issue, suggesting that initial evaluations of the IUA could adopt confirmationist biases. This does not give evaluators and users a pass from planning longer term studies, especially those related to consequences and fairness; but the initial synthesis and evaluation of the program should, at a minimum, substantiate that the development of the assessment has occurred generally as designed and that the design has empirical and theoretical support.

Dynamic Evaluation

Dynamic evaluation acknowledges that validity evaluations of most assessments are completed over a long period (Marion, 2010). Evaluators rarely have all the evidence to make conclusive judgments at a single point in time. Therefore, evaluators must engage in ongoing, dynamic evaluations as new evidence is produced. Gerasimova (2024) aptly referred to this approach as a “Living Document.” Working in this fashion requires that each proposition be written to allow judgment of whether the evidence collected thus far supports a particular claim and, further, whether alternative hypotheses are plausible. Judgment about evidence may suggest the need to collect additional evidence in support of a claim. In most operational testing programs, evaluators rarely have the luxury of concluding that the system is not working and then either ending or restarting the testing program. Rather, when critical evidence does not support the claims and intended inferences, test developers and users must quickly search for ways to improve the system.

However, the evidence might be so stacked against the intended claims that users are left only with the option of starting over.

Ryan (2002) emphasized the important role of diverse stakeholders in conducting validity studies. As we note throughout this chapter, stakeholders from diverse linguistic and cultural backgrounds must be involved in the design and evaluation of assessments to help ensure fair and equitable assessments. Therefore, the crucial activity of synthesizing validity arguments must meaningfully involve diverse stakeholders so that judgments regarding the validity of score interpretations and test use do not reflect only the dominant group.

Validity evaluators need pragmatic guidance for synthesizing the multiple lines of validity evidence so that users, developers, and stakeholders can consider this evidence, in light of the IUA, when determining the degree to which the inferences associated with the testing program are supported, with unintended negative consequences minimized. There have been some recent efforts, guided by state technical advisory committees (TACs), such as the Multi-State Alternate Assessment Consortium committee, to frame technical documentation around a validity argument (Cognia & Multi-State Alternate Assessment Consortium, 2019; see also Marion & Pellegrino, 2006). The technical report concludes with a summary of the validity evidence collected to date and its completeness. We do not suggest that this is the only way to synthesize validity evidence, and logic, but it serves as a useful example.

Multilevel and Multilayer IUA and Validity Arguments

Chen et al. (2004) discussed the need to address multiple levels in organizational research, adopting the term *multilevel construct validity*. Zumbo and colleagues (Forer & Zumbo, 2011; Hubley & Zumbo, 2011; Zumbo & Forer, 2011) expanded this concept to test score interpretation and use by reinforcing the need for evaluation of the consequences of score meaning and test use at the student, school, district, and state levels. This multilevel approach to validity investigations can better address equity and social justice issues by evaluating both individual and systemic, intended and potentially unintended negative consequences.

To illustrate how an individual test score can assume different forms and meaning, Haertel and Ho (2016) proposed a multilayer IUA that accounts for the use of various derived scores (e.g., aggregated scores) as well as the individual score. They suggested additional types of evidence to support the validity of interpretations and uses of these derived scores, based on construct definitions that differ from those specified by the test developer for the individual score. As described by Haertel and Ho (2016), common examples of derived scores in measurement include

- performance levels, such as *Proficient*, produced by standard-setting studies;
- aggregated scores, such as at the school or district level, produced by methods that change the level of measurement from the student to an aggregated group of interest;

- adjusted scores, where, for example, posttest scores are adjusted using pretest scores when measuring teacher effectiveness (see Kane, 2017); and
- score linking, for projecting from one score scale to another.

Haertel and Ho discussed how the definition of a construct takes on a different form depending on its use and how validity and fairness issues regarding test use arise from construct underrepresentation and construct-irrelevant variance. Test developers begin with the construct a test is designed to measure, but if the construct is then reframed because of its use as a derived score, evidence must be provided to support the derived score's representation of the reframed construct and its use (Haertel, 2018; Haertel & Ho, 2016). Multilevel IUA and validity arguments proposed by Zumbo and colleagues and multi-layer IUA and validity arguments proposed by Haertel and Ho allow for a more robust, comprehensive framework for evaluating score (and derived score) meaning and use.

Alternate Views of Validity Theory

Although the conceptualization of validity as pertaining to the appropriateness of test score interpretations, uses, and resulting consequences is widely recognized, some still adhere to more restricted views of validity (Borsboom et al., 2004; Cizek, 2012). For example, Borsboom et al. (2004) argued that “a test is valid for measuring an attribute if and only if: (a) the attribute exists; and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure” (p. 1061). This view tethers validity to a causal explanation, implying that differences in the tested attribute are causally related to differences in the item responses and that validation does not extend to score-use considerations. Holland (1986) illustrated the significant challenges associated with validating such causal inferences directly. Further, such a view is, of course, blind to test uses and their consequences as integral aspects of the IUA and validity argument.

Cizek (2020) offered a “comprehensive framework for defensible testing” (p. 90), in which he outlined the role of test use in determining whether test scores should be used for the intended purposes. Cizek restricted his definition of validity to inferences about variations in test scores in relation to variations in the measured construct; test use, he believes, is not part of validity. His original reasoning was that he wanted to avoid validation as “an endless process,” in Cronbach's words. Similarly, Russell (2022) proposed separating evaluation of assessment scores, decisions, and uses into investigations of (a) “instrument validity” and (b) “verifications of interpretation(s) and decision(s) made based on instrument scores. Verification explores whether the instrument scores are used appropriately to inform decisions” (p. 26). However, he did not conceptualize verification as validation.

Both Cizek (2020) and Russell (2022) required marshaling evidence in support of the intended score use, justification of this use, and an evaluation of the evidence related to the justification claims. Cizek (2020) acknowledged the significance of test use—“tests are commonly used to inform important decisions” (p. 16)—but then argued for the separation of score inferences and uses. We contend that separating inferences

from uses, where only inferences are validated, and then bringing them back again later in the process risks having certain critical validation (e.g., uses and consequences) fall by the wayside.

Theories of Action and Validation

Haertel (2018) noted that an IUA is like a “theory of action that explains how testing is supposed to confer a benefit in some particular situation” (p. 207). Other researchers have suggested using a theory of action more directly to frame an IUA and validity argument. Ryan (2002) and others suggested that providing a theory of action is a useful starting point for users when developing a more complete IUA. The theory of action requires the explication of the intended components of an assessment and decision system, as well as the mechanisms by which a test user could reasonably expect to get from one step to the next. To address equity issues, Kane (2006) and Shepard (1993) encouraged evaluators to attend to values and consequences when evaluating a decision procedure, especially when a testing program is used as a policy instrument (as in state tests). When conducting such a validity evaluation, the values inherent in the testing program must be made explicit, and both intended and unintended consequences of decisions based on test scores must be evaluated. Values favoring social justice aims must be prioritized when considering the consequences of a testing program. Theories of action have been applied to many types of assessments, including assessment systems (Bennett, 2010; Bennett et al., 2011), assessments for students with significant disabilities (Marion, 2010; Marion & Perie, 2009), language testing (Bachman, 2002), and certification testing (Crisp, 2017).

The Cognitively Based Assessment of, for, and as Learning (CBAL) program was guided by an extensive theory of action (Bennett, 2010). The CBAL theory of action explicates the processes and mechanisms connecting the design of the assessment program with the intended outcomes. It emphasizes “the effects of the assessment system on individuals and institutions as well as to the underlying mechanisms behind those effects” (Bennett, 2010, p. 71). In the case of CBAL, these mechanisms focus on such considerations as how well teachers use tools and resources, including the use of research-based learning progressions, to adjust instruction to improve student learning.

Assessments for students with significant cognitive disabilities are relatively new components of state and district testing systems, initially required as a result of the Individuals With Disabilities Act and the No Child Left Behind Act in the early 2000s. Researchers, test developers, and state specialists grappled with how best to assess the KSAs of students who had difficulty communicating their thinking. One advantage of working in this area is that, because many of these assessments were being designed for the first time, researchers were able to articulate a validity evaluation plan as part of the development work (e.g., Marion & Pellegrino, 2006).

Given the field’s lack of clarity around these alternate assessments, a well-developed theory of action is perhaps even more critical than it might be for other validation

initiatives. Policy makers, developers, stakeholders, and technicians need to explicate why implementing an alternate assessment for students with the most significant cognitive disabilities would lead to improved educational opportunities for these students. In addition to the *why*, these theories of action, like the CBAL case, described the mechanisms by which these improved learning opportunities would occur. For example, one might postulate that alternate assessment scores provide accurate depictions of what eligible students know and can do so that teachers can offer more appropriate learning opportunities for them. The evaluator and user must specify the mechanism by which these score reports will lead to anticipated changes in teaching practices, such as targeted instruction or more appropriate curricular materials (Marion, 2010; Marion & Perie, 2009).

Bachman (2002) provided an example of an important use of a theory of action when developing a validity argument. He used theories of action as a bridge to a validity argument to illustrate how to attend to potential alternative explanations for score meanings and unintended negative consequences. He noted that performance-based assessments were designed to measure complex constructs, such as those found in language testing. While the validity issues associated with performance assessments have received considerable attention (e.g., Lane & Iwatani, 2016; Lane & Stone, 2006; Linn et al., 1991; Shepard, 1993), Bachman's theory of action helped uncover alternative interpretations of test scores due to language knowledge, content knowledge, and their interaction.

Theories of action provide a framework for bringing together meaning and impact claims, with an emphasis on claims about the intended consequences of the assessment as well as potential unintended negative consequences for both individuals and groups (Bennett, 2010; Marion & Perie, 2009). This has become increasingly important over the past 20 years as tests have played a more prominent role in educational and social policy, resulting in unintended negative consequences for marginalized students. Haertel (2018) discussed two classes of indirect effects of testing: incentive effects and messaging effects. He argued that the use of testing for its indirect effects is a significant change: "Instead of using tests to *evaluate* educational treatments, testing itself had now *become* the treatment" (p. 207, emphasis in original). For example, an incentive effect occurs when the teacher tells students, "This will be on the test." The students have an incentive to pay attention and study. Messaging effects often occur through the news media and shape perceptions of schools and schooling, particularly when tests are part of accountability systems.

The use of a theory of action to frame the IUA serves to identify the particular claims about benefits that are most critical to investigate (Shepard, 2016), as well as potential unintended negative consequences, particularly for historically underserved students. The development of a theory of action, with input from multiple stakeholder groups, fosters the consideration of different values and social justice concerns (Ryan, 2002). Finally, evaluating the IUA through the validity argument is not a mere formula-driven exercise. The many pieces of evidence are just that; they do not form a validity argument in and of themselves (Haertel, 1999). The various instances of theoretical, logical, and empirical evidence need

to be woven into a coherent argument. Creating and appraising the theory of action is one way to help synthesize sources of validity evidence in the evaluation of the IUA.

CONSEQUENCES IN VALIDITY THEORY AND VALIDATION

The best way to encourage adequate consideration of major intended positive effects and plausible unintended negative effects of test use is to recognize the evaluation of such effects as a central aspect of test validation. (Linn, 1997, p. 16)

The soundness of test-based decisions and uses has been an integral aspect of validity, resulting in the need to identify the IUA's claims regarding intended positive consequences and potential negative consequences (Cronbach, 1971; Kane, 1992, 2002, 2006; Lane et al., 1998; Linn, 1997, 2009; Shepard, 1997). The validity argument should attend to potential adverse consequences, particularly for individuals who have been underserved by the educational system. Five decades ago, Cronbach (1971) considered the evaluation of decisions and actions based on tests as part of validity evaluation. In response to the debate about whether consequences of test use are an aspect of validity, Linn (1997) argued that "the evaluation of consequences rightly belongs in the domain of validity" (p. 16) and, thus, called for the appraisal of intended positive consequences and potential unintended negative consequences as fundamental to validation efforts. Popham (1997) and Mehrens (1997) argued in opposition.

Messick's (1989) progressive matrix made the role of consequences explicit in validity theory. The inclusion of the appraisal of the soundness of decisions based on test scores underscores the need to examine consequences as a "logical part of the evaluation of test use, which has been an accepted focus of validity for several decades" (Shepard, 1997, p. 5). We contend that score meaning cannot be evaluated without considering use and resulting consequences; therefore, we disagree with the contention that validity pertains to only score meaning (e.g., Cizek, 2020).

Because consequences are integral to validity, IUAs should identify claims associated with intended positive consequences and potential unintended negative consequences of the testing program. An examination of consequences in the validity argument involves a critical evaluation of both the anticipated positive and the potential unintended negative consequences of test use (Kane, 1992, 2006; Lane, 2014; Lane & Stone, 2002; Lane et al., 1998; Linn, 1997; Shepard, 1993, 1997). Judicious attention to potential adverse consequences, particularly for historically underserved individuals, is critical as we work toward a more equitable and just society in which the negative effects of testing are minimized. Adverse social and personal consequences of testing bear on both validity and fairness issues. As articulated by Campbell (1979), "the more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 49).

Tests as mechanisms of social action have direct and indirect consequences. In consideration of validation as a persuasive argument, Cronbach (1988) asserted that "*the argument*

must link concepts, evidence, social and personal consequences, and values” (p. 4). Although consequences of test use have been included in the evaluation of test decision procedures, the attention to social consequences, differential impact, and adverse systemic effects for different groups did not occur until the 1960s. The civil rights movement and corresponding legislation elevated discussions of the social consequences of testing, with a focus on the adverse effects on ethnic and racial minorities and women (Cronbach, 1971, 1988; Messick, 1965, 1975, 1980). Social consequences were given prominence because of their impact on society, not just those involved in the testing process (Kane, 2016). This resulted in the need to integrate such consequences into validity and validation.

Messick (1989) explicitly addressed the relationship between validity and consequences: “Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test score and other modes of assessment” (p. 13). Newton (2012) contended that justifying test use is an enormous undertaking, believing it is reasonable “to separate distinguishable elements of evaluation into manageable components—validity, efficacy, legality, and so on” (p. 15), and suggested the need for cleaner distinctions when consequences are part of validity arguments. Cizek (2020) argued for decomposing the components of a validity evaluation, especially when consequences are part of the mix. But that begs the critical question of synthesizing the evidence and logic from the various components into some sort of judgment regarding the validity of using the test for a particular purpose. Newton and Shaw (2014) addressed the challenges of synthesizing the evaluation components within a cost–benefit judgment concerning the acceptability of a testing policy.

Kane (2006) supported Cronbach’s (1971, 1988) position that consequences have an integral role in the validity argument and that “tests that impinge on the rights and life changes of individuals are inherently disputable” (Cronbach, 1988, p. 6). Those responsible for providing evidence in support of the validity argument have an obligation to evaluate whether a use has “appropriate consequences for individuals and institutions, and especially to argue against adverse consequences” (Cronbach, 1988, p. 6). Kane (2013) later wrote that consequences that have “potential for substantial impact in the population of interest, particularly adverse impact and systemic consequences” (p. 61), should be a central aspect of the validity argument. The evaluation of consequences of testing programs is an inherent aspect of the IUA and validity argument.

Consequences and the 2014 *Standards for Educational and Psychological Testing*

The 2014 *Standards* backtrack on the connection between scores and uses by conceptualizing validity as “the degree to which evidence and theory support the interpretation of test scores *for* [emphasis added] proposed uses of tests” (AERA et al., 2014, p. 11). This distinction is critical: It suggests that use alone is not central to validity. The 2014 *Standards* also adopted Messick’s limited view that consequences are considered an aspect of validity if they result from other interpretative threats to validity—principally, construct irrelevance or construct underrepresentation.

An evaluation of subtle, but critical, modifications in the 2014 *Standards* validity chapter uncovers how consequences play a marginal role in validity theory when compared with the 1999 *Standards* (Lane & Niepokoj, 2018). The comment associated with Standard 1.6 in the 2014 validity chapter states that

certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices. . . . To the extent that such claims enter into the justification for a testing program, they become part of the argument for test use. Evidence for such claims should be examined—in conjunction with evidence about the validity of intended test score interpretation and evidence about unintended negative consequences of test use. (AERA et al., 2014, p. 24)

In contrast, the 1999 *Standards* clearly placed consequence in the validity argument, arguing that “to the extent that such claims enter into the justification for a testing program, they become *part of the validity argument for test use* [emphasis added] and so should be examined as part of the validation effort” (AERA et al., 1999, p. 13).

Note how the 2014 *Standards* replaced “validity argument for test use” with “argument for test use,” implying that the 2014 standard is not part of a validity argument. The 2014 *Standards* similarly removed consequences from the validity argument for tests used as levers of educational change.

Two other chapters in the 2014 *Standards* addressed consequences. Standard 12.1 in the “Educational Testing and Assessment” chapter supports the need to consider consequences for educational assessments, but it does not acknowledge that consequences are part of the validity argument:

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described by those who mandate tests. It is also the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences as feasible. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test developer and/or user. (AERA et al., 2014, p. 195)

The comment for Standard 12.1 refers to the validity of score interpretations—not score uses, test uses, and the resulting consequences:

The need to monitor the impact of educational testing programs relates directly to fairness in testing, which requires ensuring that scores on a given test reflect the same construct and have essentially the same meaning for all individuals in the intended test-taker population. (p. 195)

Standard 13.7, in “The Uses of Tests for Program Evaluation, Policy Studies, and Accountability” chapter, refers to consequences:

When tests are selected for use in evaluation or accountability settings, the ways in which the test results are intended to be used, and the consequences they are

expected to promote, should be clearly described, along with cautions against inappropriate uses. (p. 212)

The only reference to consequences as potentially included in the validity argument surfaced in this comment for Standard 13.7:

The use of tests is often justified on the grounds that it will improve the quality of education by providing useful information to decision makers and by creating incentives to promote better performance by educators and students. These kinds of claims should be made explicitly when the system is mandated or adopted, and evidence to support their validity should be provided when available. The collection and reporting of evidence for a particular validity claim should be incorporated into the program design. (p. 212)

The 2014 *Standards* endorsed the importance of evaluating the consequences of test use but, until page 212, did not consider consequences as part of the validity argument. Throughout the validity chapter and elsewhere in the *Standards*, the validity of score interpretations *for* proposed uses was adopted instead of the validity of score interpretations *and* uses. This seemingly subtle difference arguably limits the role of consequences as an integral aspect of validity. Newton and Shaw (2016a) discussed the 1999 and 2014 *Standards* references to validity as related to score interpretations for proposed uses, instead of “and” uses, calling for clarification of the treatment of both use and consequences in validity theory. The 2014 *Standards* represented an unfortunate backtracking by understating the importance of consequences in validity and validation.

Consequences and Credentialing Testing Programs

Tests that certify or license individuals for professional practice must provide a definition of minimally competent qualified candidates (see Margolis et al., this volume). Such definitions rely on the judgments of content experts. Kane (2004) identified pertinent inferences for a credentialing testing program’s IUA, including the decision inference. The decision inference is that those scoring above the cut score have the requisite KSAs for entry-level practitioners, and those scoring below the cut score lack the necessary KSAs and pose a risk to the public. Both intended and unintended consequences arise from this decision inference. The practice analysis that identifies the critical KSAs in the domain, the test design and development process, and standard setting provide backing to support this decision inference—all of which rely on expert judgments.

Consequences and Setting Performance Standards

Factors that introduce uncertainty in the cut score include the standard-setting panel size and composition, the standard-setting method, the implementation of the standard-setting method, and the types of data and information provided to panelists to make informed judgments (Hambleton & Pitoniak, 2006; see also Ferrara et al., this volume). Various factors can lead to inconsistencies in a standard-setting method, such

as regression effects in item ratings and the tendency for statistically easy items to be rated as more difficult and statistically harder items to be rated as easier (Shepard et al., 1993; Wyse, 2018). In the bookmark method, for example, the choice of response probability value to be used for anchoring items to the score scale, response probability 50 versus response probability 67, will yield different cut scores (Wyse, 2011). Competing values, beliefs, and assumptions inherent in any standard-setting method pose trade-offs. Standard setting, and the choices made throughout the standard-setting process, can impact individuals' livelihood.

Kane (1994) identified two inherent assumptions in his discussion of arbitrariness in setting performance standards. First, candidates who score above the cut score meet the performance standard, and those scoring below do not. Second, the standard is reasonable for the purpose of the decision. This second assumption rests on values and beliefs about consequences in an adopted policy that are "likely to be subject to dispute" (p. 426). These policy decisions are consequential.

Shepard (1980) expressed her concern with the dichotomization of the scale for decision-making: "Individuals immediately on either side of the standard will be virtually indistinguishable from one another. . . . Pass-fail distinctions near the cutoff will have poor validity because a continuum of performance has been 'arbitrarily' dichotomized" (p. 448). In the context of teacher certification tests, Jaeger (1990) echoed Shepard's concern:

No conventional test validation procedure will provide evidence that any score-based dichotomization of the ability scale into two categories labeled "competent" and "incompetent" is correct. We know that dichotomization is judgmentally based, arbitrary, and, wherever placed on the ability scale, will not result in reliable differences in distributions of performance on any valued criterion for groups adjacent to the point of dichotomization. (p. 18)

In a similar vein, Kane (1994) stated that a defensible standard is chosen from a range of performance standards, and there is arbitrariness in this choice insofar as there is no technical basis for identifying a particular "correct" standard (p. 434). Various stakeholders, such as content experts and policy makers, may have different assumptions and values they apply to standard setting. While advances in standard setting have minimized the degree of arbitrariness, different standard-setting methods and different standard-setting panels can produce different results, which threatens the generalizability of the results across different standard-setting methods, panels of experts, contexts, and occasions.

These issues with standard setting and resulting cut scores (or performance standards) apply to any testing program that sets cut scores for decision-making. We acknowledge that society requires certain cut scores (e.g., competent to earn a driver's license) even though, as noted previously, adjacent candidates separated only by a point or two are generally indistinguishable. This is why setting cut scores requires consequential evidence. In the context of testing for college and career readiness, for example, Zwick (2018)

cautioned that dichotomization into readiness levels reduces accuracy of prediction while “paradoxically convey[ing] an increased sense of certainty: the message is that a ‘college-ready’ student is qualitatively different from an ‘unready’ one, even though their scores may differ by a psychometrically negligible amount” (p. 101). Using additional measures that account for variance in the criterion, rather than relying on a single score, would improve classification accuracy (Zwick, 2018).

Maximizing Equity in Standard-Setting Policy Decisions

Policy decisions that lead to the performance standards and corresponding cut scores reflect values associated with the potential consequences. The justification of a decision rule depends on an evaluation of intended positive and potential negative consequences of a decision rule. Cronbach (1989) argued that test-based decision rules could reflect value assumptions aiming to increase diversity. Policy makers balance competing values when evaluating proposed cut scores. Any differential impact for test takers from different cultural and language groups requires careful consideration in decision-making. Information on measurement and sampling error, along with a reasonable range for the cut score, informs the policy makers’ decisions on the final cut score. There are different consequences for the two types of decision error: not passing a test taker who nonetheless has the requisite KSAs (false negatives) and passing a test taker who lacks the requisite KSAs (false positives). Consideration of the consequences of each type of error is needed when setting the cut score (Hambleton & Pitoniak, 2006). Information on differential impact on population groups provides the opportunity for decision-makers to maximize diversity while maintaining rigorous standards. All involved in the standard-setting process, whether for credentialing testing or for K–12 achievement testing, should have a clear understanding of the judgment and uncertainty inherent in setting cut scores and the potential effect of decision rules on the livelihoods of individuals.

Legal issues that arise in credentialing testing programs bear on constitutional claims of due process and fairness, where fairness necessitates that the testing process avoids arbitrary procedures (Atkinson, 2012). The use of testing as a social control mechanism, and the deprivation of rights without due process, unduly affects historically underserved groups. The adherence to professional standards and best practices in testing promotes thoughtful decisions in testing practices (including standard setting) while enhancing informed judgments. That said, a degree of arbitrariness will always remain. Therefore, an independent evaluation of standard setting, along with other testing practices, is recommended.

Consequences When Extrapolating to Performance in Practice

The backing for the decision and extrapolation inferences for credentialing tests also includes evidence from the test design and development process, such as evidence from the practice analysis that identifies the KSAs necessary for effective practice. Kane (2009) argued that the test is “viewed as a measure of KSAs that are considered necessary but not sufficient for effective practice, and as a result, the test scores are not expected to provide

very accurate estimates of performance in practice [emphasis added] for individual candidates” (Kane, 2009, p. 142). What is assessed is not sufficient for effective performance; when test items refer to a practical scenario, the testing situation differs substantially from the practice situation, which affects the validity of extrapolation from the score to the application of the knowledge and skills in practice (Kane et al., 2017). The behaviors and processes involved in responding to multiple-choice items are unlike those when responding in practice, so there is a need for additional backing that shows the relationship between performance on multiple-choice items and performance on high-fidelity simulations (Haertel, 2004). This is reinforced in the 2014 *Standards*:

When the responses elicited by a test neither sample nor closely simulate the use of tested knowledge or skills in the actual criterion domain, participants are not likely to approach the task with such clear understandings of adequacy or quality. (AERA et al., 2014, p. 108)

This jeopardizes the validity of inferences involving extrapolation to both the broader domain of KSAs and the practice domain. Adverse consequences can arise if there are concerns with any of the design and measurement procedures that affect the validity of inferences and uses. There are limitations to the validity of inferences and uses for credentialing testing programs that rely on selected-response item types, with or without a practice scenario. Even if one adopts the narrow view of consequential evidence, construct underrepresentation and construct-irrelevant variance can differentially affect diverse groups or individuals, resulting in validity and social justice issues. This is one of the rationales for the use of performance assessments: They arguably represent the intended practice with more fidelity compared with items that only serve as a proxy at best.

Consequences and K–12 Assessment and Accountability Systems

The intent of current test-based accountability systems is to improve the educational opportunities afforded to all students by holding schools, educators, and (sometimes) students accountable (see Ho & Polikoff, this volume). As educational change agents, tests serve as powerful tools to support educational reforms while at the same time exposing societal and educational inequities. Therefore, the appraisal of test-based decisions and uses in terms of their consequences is integral to the validation efforts of tests used in these systems (Lane, 2014, 2020). The use of a test as an instrument of policy requires an evaluation of the consequences of its use (Cronbach, 1982).

Educational reform policies that involve testing are intended to result in certain consequences, including improved student achievement and learning, enhanced instruction, narrowed achievement gaps, and students being prepared for college and careers. Tests used in accountability systems provide policy makers with tools to hold schools, administrators, educators, and students accountable for student achievement and learning. Tests appeal to policy makers as agents of educational change because they are relatively inexpensive, can be mandated externally and implemented relatively quickly, and have quantifiable results that are perceived to be objective or at least independent (Linn, 2000). Advocates view

test-based accountability systems as a means to raise academic standards for students and to focus instruction on content aligned with rigorous content standards. The intent of tests as educational change agents involves incentivizing and/or signaling to educators and leaders to provide all students an opportunity to learn rigorous content described by the standards and to achieve the knowledge and skills reflected in those standards.

Evaluating the match between the intended interpretation and uses of test scores and the enacted interpretations and uses of test scores is integral to the validity argument for test-based accountability systems. The validity of a decision inference is threatened if there is a mismatch between intended and enacted interpretations and uses, resulting in unintended social and personal consequences. Enacted interpretations and uses by educators and administrators are more diverse and complex than intended interpretations and uses (Coburn & Turner, 2011; Moss, 2013, 2016), and they may result in undesirable consequences for some individuals or groups. Tests are used locally in different contexts by educators and administrators to address a variety of their own needs (e.g., grade promotion), and these local uses need to be understood and evaluated in terms of their appropriateness and potential negative consequences for students.

Decisions made in test design, development, and use rest on value judgments that affect the consequences of test use (Kane, 2006; Messick, 1989). Value judgments are involved in the articulation of the intended purposes and uses of a test, the specification of the construct (or content standards) to be assessed, the design and development of items and scoring procedures, the development of performance-level descriptions, the setting of cut scores for determining student proficiency or passing, and the design of score reports and accompanying interpretive guides. Value judgments inherent in each of these test design decisions impact the validity of test score interpretations, uses, and resulting consequences. Value judgments are also inherent in the specification of the intended positive consequences and in identifying potentially unintended adverse consequences.

Different stakeholders—policy makers, advocacy groups, administrators, educators, parents, students, business leaders, the community—have different, and sometimes conflicting, perspectives on the purposes and uses of tests in instruction and accountability systems because of corresponding differences in their assumptions and values. One way to address these potentially competing interests is to construct different IUAs (Bachman, 2002), although such an approach would make validation an even more cumbersome endeavor. Nevertheless, the documentation of competing assumptions and values among stakeholder groups will promote a better understanding of the potential consequences, both positive and negative, of test use. It would be more efficient, however, to do so within a single IUA.

The use of tests as policy levers bears on equity and social justice issues. The use of test-based accountability systems typically has different consequences for different student groups, schools, and districts, which should be examined as part of the IUA and validity argument (Lane & Stone, 2002). Differential test outcomes for students living in economically marginalized communities, African American students, Hispanic students, Indigenous students, English learners, and students with disabilities have led

to undesirable and inequitable consequences, such as a narrowing of the curriculum and instruction (Booher-Jennings, 2005)—ironically, for the very same students the tests were intended to support. The evaluation of test use in accountability systems should determine whether the intended positive consequences are realized and potential unintended negative consequences are minimized. The inequities of educational opportunities compromise the validity and fairness of test score interpretations and uses, as well as the use of tests as vehicles for educational change, all resulting in undesirable consequences.

Validity Evaluation for Consequences in K–12 Contexts

Evaluating the effectiveness of tests used as policy instruments in accountability systems requires attention to how well the tests are achieving their goals, both in terms of the motivation and practices of school administrators and educators and in terms of student achievement and learning (Lane & Stone, 2002). Those who mandate and use tests are obliged first to make a case for the appropriateness of the decision rules in the use context and then to consider whether any potential unintended negative consequences outweigh intended positive consequences. The articulation of intended and potential unintended consequences should occur in conjunction with the specification of the test purposes and claims. The validity argument should entail a systematic evaluation of whether, and how, the test achieves its goals, as well as the costs of doing so. The relevant stakeholders, who may hold different, and perhaps competing, values, should be involved in identifying the questions for evaluating both positive and potentially negative consequences of test use.

A principled approach to the evaluation of consequences must be central to the validity argument to avoid continued “begging the question” (Kane, 2006). The claim that test-based accountability systems will lead to better instruction and student achievement often is taken for granted. Rather, integral to the IUA and validity argument is the empirical evaluation of whether testing programs, used as levers for educational change, have the intended positive effects on instruction and student learning while minimizing potentially negative unintended effects.

Lane and colleagues developed and implemented a conceptual framework for evaluating the consequences of assessment and accountability systems—intended and unintended, positive and negative—for a state performance-based assessment program that held schools and educators accountable for improved instruction and student learning (Lane & Stone, 2002; Lane et al., 1998). The proposed IUA addressed the intended consequences at multiple levels (individual, school, district, state) by articulating a set of claims that could be either supported or refuted by logical and empirical validity evidence at each level. This involved specifying the needed evidence relevant to both positive and potentially negative outcomes for different stakeholders (students, teachers, administrators) and different levels (state, district, school, class, student). For each form of evidence, the sources to provide the evidence were delineated. The framework stipulated the need for contextual information to inform the evaluation of consequences, such as the percentage of students receiving free or reduced lunch; amount of funding per student; access to quality curriculum and materials; and stability of administrators,

teachers, and students. Contextual information also can identify barriers to achieving intended outcomes that must be overcome. The framework provides a mechanism for synthesizing the evidence for consequences in the validity argument.

Consequences and the Use of Performance Assessments in Education

Consequential evidence in support of the validity argument is “especially compelling for performance-based assessments . . . because particular intended consequences are an explicit part of the assessment system’s rationale” (Linn, 1993, p. 6). If a performance assessment is intended to achieve positive outcomes, the enacted program should achieve these goals while minimizing negative consequences. The implementation of large-scale performance assessments in the 1990s improved instruction and student learning, with a greater emphasis on problem-solving, critical thinking, and reasoning (NRC, 2010). Lane and colleagues examined validity evidence for the consequences of the Maryland School Performance Assessment Program (MSPAP), which served as a lever for educational change. To provide consequential validity evidence, they used growth model analyses to examine the impact of the performance assessment and accountability system on both student learning and classroom instruction (Lane et al., 2002; Parke & Lane, 2008; Parke et al., 2006; C. A. Stone & Lane, 2003). Teacher-reported use of performance-based instruction accounted for differences in school performance on MSPAP reading, writing, mathematics, and science. Schools emphasizing performance-based instruction, such as engaging students in critical thinking and reasoning, had higher MSPAP scores than schools where instruction was less performance based (Lane et al., 2002; C. A. Stone & Lane, 2003). A systematic evaluation of mathematics classroom instructional artifacts revealed a stronger alignment with MSPAP and the state content standards for higher achieving and higher gain schools compared with lower achieving and lower gain schools (Parke & Lane, 2008).

Furthermore, trends in mathematics on the state National Assessment of Educational Progress and MSPAP mathematics assessments were similar, suggesting that improved student performance on MSPAP most likely was real (Linn et al., 2002). These results support extrapolating score-gain inferences on MSPAP to other assessments measuring somewhat different content and using different item formats. Such positive results, in part, may have resulted from schools using MSPAP data, along with other information, to guide instructional planning (Ferrara, 2010; Michaels & Ferrara, 1999).

Contextual information can inform inferences and resulting actions when using test scores to evaluate the quality of education (Haertel, 1999). In the MSPAP studies, for example, the percentage of students in a school who receive free or reduced-price lunch (a school-level proxy for socioeconomic status [SES]), was significantly related to school-level MSPAP performance (Lane et al., 2002; C. A. Stone & Lane, 2003). However, there was no significant relationship between SES and school performance *gains* in mathematics, writing, science, and social studies. This offers some evidence that MSPAP did not have an adverse impact on students who lived in economically disadvantaged areas, given the

comparable growth in school scores in these subjects regardless of school SES. Stakeholders might argue that greater gains for lower SES schools should have occurred insofar as initial performance was lower for these schools. In short, comparable gains in performance may not be a sufficient positive consequence since the goal is to decrease achievement gaps.

Evaluating the consequences of performance assessments provides evidence primarily to support the decision/use inference. Evidence is also needed in support of the performance, scoring, generalization, and extrapolation inferences. In particular, the generalization inference requires careful scrutiny, given that performance assessments typically comprise a small number of tasks.

Consequences and the Use of College Admission Tests as High School State Assessments

The IUA and validity argument must consider consequences when states use college admissions tests as high school state assessments for their accountability systems (see Camara et al., this volume, for a discussion of tests used for admissions). Such tests often underrepresent the intended construct by measuring only a portion of state content standards, which may tempt teachers to either (a) narrow the curriculum to cover only tested material or (b) maintain fidelity to the intended curriculum and potentially leave their students underprepared for the high-stakes test. Either choice may carry significant negative consequences for students. The validity argument must include an independent evaluation of the alignment between the admissions test and the state content standards. There are gaps in the alignment between content standards and admissions tests (Achieve, 2018), which, for states that adopt college entrance exams for accountability, could change how mathematics and English language arts are taught. This potential negative consequence should be weighed against the positive consequence of using a college entrance exam as the high school accountability test, such as furthering the educational opportunities of students who may not have considered higher education. However, validity threats associated with college entrance exams cannot be considered in the absolute. The evaluation must be weighed against the threats of using a state survey test, such as lack of motivation and a weak link to students' course-taking patterns (Marion, 2018b; Marion & Domaleski, 2019). States that have adopted college entrance tests should address these consequences in their IUA and validity argument by providing evidence related to content alignment, changes in college enrollment rates for students (especially from underserved populations), and the impact on instruction at the classroom, school, and district levels.

Maine was one of the first states to adopt a college entrance exam as its state test in high school. Maine began administering the SAT in 2006 as its high school accountability test, but needed to augment the SAT mathematics section with additional items to improve its alignment with state content standards. Hurwitz et al. (2015) found that a 2%–3% increase in college-going rate was associated with the use of the SAT in Maine. Michigan adopted the ACT as its high school accountability test, which resulted in approximately 50% more low-SES students achieving a college-ready score (because most of those students would not have tested previously). Unfortunately, this translated into an increase of only 6% more of these low-SES students later enrolling at a 4-year

institution (Hyman, 2017). Nonetheless, these studies provide some backing for the use of college admission tests to increase higher education opportunities for students, including those who are underserved.

Consequences and Language Testing

Researchers in various academic disciplines have given test use and resulting consequences a prominent role in validity theory and validation. In the context of language assessments typically used for selection purposes, Bachman (2005) and Bachman and Palmer (2010) provided an assessment use argument (AUA), which agrees in spirit with Kane's IUA. Their AUA offers a framework for linking test performance to decisions and use, delineating the intended consequences and potential unintended adverse consequences, as well as the corresponding validity argument. Figure 4.1 displays a simplified version of this AUA. The AUA links score interpretations to decisions and consequences, specifying four claims about outcomes: intended and actual consequences, intended and actual decisions, intended and actual interpretations, and test scores.

The articulation of the intended beneficial, and potentially unintended, consequences of test use occurs at the beginning of the testing process. The warrant associated with consequences in the AUA states that “the *consequences* of using the assessment and making intended decisions will be beneficial to individuals, the program, company, institution, or system, or to society at large” (Bachman, 2005, p. 19). This sentiment captures Messick's

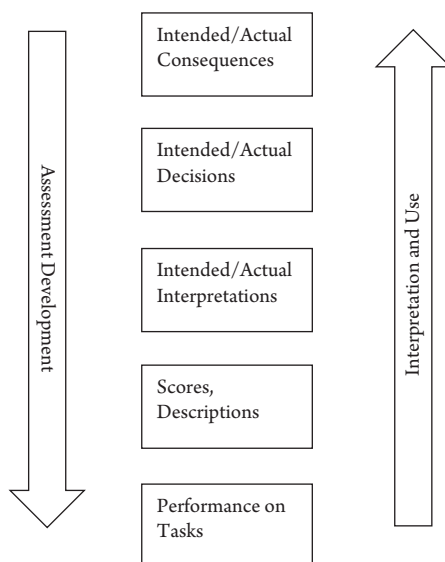


FIGURE 4.1
Inferential Links in the Assessment Use Argument

Note. Adapted from “Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World,” by L. F. Bachman and A. S. Palmer, 2010, Figure 5.1, p. 91. Oxford University Press.

notion of social consequences embedded in his progressive matrix (Table 4.1). But in contrast to Messick's position that consequences are an aspect of validity only if they can be traced to construct underrepresentation or construct-irrelevant barriers, Bachman (2005) believed that "it is quite possible for adverse consequences and inappropriate uses of tests to occur that are *not* a result simply of sources of invalidity" (p. 16) in the score interpretation.

Importantly, in addition to emphasizing intended score interpretations, decisions, and consequences, the AUA also emphasizes *actual* interpretations, decisions, and consequences. This broadens the evaluation demands by now including the appraisal of actual test use and resulting consequences. The inclusion of consequences as an integral aspect of the AUA allowed for encompassing investigations of test consequences on teaching and learning as part of the AUA. Lack of access to the assessment because of financial, geographic, personal, and educational limitations ultimately restricts individuals from consideration in the decision-making process (Bachman, 2005). The claims and counterclaims specified in the AUA provide a framework for guiding assessment design and developing procedures for appraisal of the test use and resulting consequences (Bachman, 2005).

Chapelle et al. (2007, 2010) used aspects of Kane's IUA and Bachman and Palmer's AUA in the context of an assessment of English as a foreign language. In doing so, they provided an example of a decision/use inference⁴ for Test of English as a Foreign Language (TOEFL) scores used for admission decisions at English-medium universities. The warrant for the use inference is that these scores are useful for making decisions about admissions, placements, and curricula for test takers. Using TOEFL scores for admissions and placements assumes there will be appropriate score interpretations by admission officers, test takers, and teachers, while using these scores for curricula decisions assumes the test will have positive consequences for English language instruction. Support for this warrant and the corresponding assumptions comes from standard-setting studies, score reports, interpretative materials, available instructional supports, and studies examining the consequences for instruction (Chapelle et al., 2007, 2010). Counterevidence includes any potential differential impact for individuals from different cultural and language backgrounds. Additional backing for placement decisions would include evidence that the scores appropriately identify the needed level of English instruction for individuals.

Investigations of Consequences of Selection Decisions

The validity argument for using language tests for making selection decisions should include the appraisal of the relationship between language test scores and subsequent performance in college. This relationship has been evaluated in terms of correlation coefficients, but this approach has limitations. Cho and Bridgemen (2012) reported that although the correlation between the TOEFL Internet-Based Test (iBT) and GPA is relatively small, the likelihood of being in the top 25th percentile GPA group doubled for students who had iBT scores in the top 25% compared with those who had scores in the bottom 25%. They discussed several typical factors that affect correlation coefficients for selection test scores, including restriction of range for the test scores as a result of

only being able to use admitted students and lack of variability in university GPAs. Deygers et al. (2018) investigated whether a small sample of international students at Flemish universities had the language skills needed for successful performance in college even though they failed the language tests. The researchers posited that using the tests as gatekeepers to universities for admission decisions is just, with the warrant that students who do not pass would not meet the linguistic demands of college and hence would be unsuccessful. Their results did not support this claim and warrant because one in five successful university students would have been denied admission had their test scores been used for the admission decision. Although there also were false positives—students who passed the tests but subsequently demonstrated significant language difficulties at the university—Deygers et al. argued for paying attention to the false negatives that result in restriction of opportunities (Rawls, 2001).

Investigations of Consequences on Instruction and Learning

In the context of the testing English as a foreign language, Alderson and Wall (1993) posited a series of potential consequences of language tests: what and how teachers teach, what and how learners learn, the rate and sequence of teaching and learning, the degree and depth of teaching and learning, and teachers' attitudes toward content and method of teaching and learning—all of which could be part of the IUA and validity argument. Similar to earlier studies (Alderson & Hamp-Lyons, 1996; Alderson & Wall, 1993), Barnes (2016) concluded that test preparation materials can improve teaching and learning in English language preparation classes for the TOEFL iBT, which provides an integrated approach to assessing speaking, listening, reading, and writing. Barnes (2016) found that the TOEFL iBT preparation materials determined the content of instruction (including speaking) and, to some extent, the way this content was taught. However, the curriculum arguably should not be defined solely by test preparation materials.

Concluding Thoughts

Test use and resulting consequences are inseparable from validity. Therefore, IUAs should identify claims associated with intended positive consequences of the testing program as well as potential unintended negative consequences. Validation efforts should pay close attention to potential adverse social and personal consequences, particularly for individuals who have been historically underserved, because of the impact consequences have on society and the livelihood of individuals.

ASSESSMENT OF INDIVIDUALS WITH VARYING CULTURAL AND LINGUISTIC BACKGROUNDS

It is imperative that we take the next step, which is to begin empirically based investigations in the design, validation, and meaningful utility of culturally responsive assessments (Hood, 1998, p. 4).

There has been a strong call for more equitable and culturally inclusive testing practices and uses so that one cultural group is not privileged over another. This call stems from the increasing diversity of tested groups as well as the recognition of structural racism and the role testing plays in it (Gould, 1981; Randall, 2021; see also the following chapters in this volume: Ercikan & Solano-Flores, Rodriguez & Thurlow, and Zwick). Culture is a mediator of human practices, values, and beliefs (Cole, 1996). Testing is a cultural practice situated within a cultural context with tests as cultural artifacts, reflecting primarily the ideologies, values, practices, and beliefs of the culture of those who mandate and design the tests. Importantly, the field is moving toward a more inclusive testing paradigm, allowing for the design of assessments that provide more valid and equitable score interpretations and uses for culturally and linguistically diverse groups.

Shifting the current testing paradigm requires adopting a critical lens to examine what is valued, and not valued, in the testing process and by whom—from conceptualizing the construct to articulating the intended and potential negative consequences of test use. This lens entails critical analyses of our own culturally informed values, beliefs, and assumptions to uncover potential biases underlying the testing process. Refocusing testing to address validity and equity issues for diverse cultural and linguistic groups requires the inclusion of critical individuals from these groups who can advocate for such identities in the design, development, implementation, use, and validation of tests. Varying, and perhaps competing, perspectives and values need to be considered throughout the testing process. This view of testing has significant implications for validity theory and practice, particularly by considering the linguistic and cultural practices that influence how students learn, as well as the learning theories that explain how students are expected to learn within certain domains.

Although current practices, such as item sensitivity and bias reviews and examination of DIF, have improved the assessment of individuals from diverse backgrounds in some cases, these practices are inadequate. Rather, the field needs to develop and implement testing practices and methodologies that produce more equitable and fair assessments, with a corresponding program of research to evaluate whether these assessments are, indeed, more equitable.

We first discuss testing as a situated cultural practice before considering the importance of learning theories in shaping the design and validation of more equitable educational assessments. We discuss the challenges of framing and conducting validity evaluations of assessments designed to be responsive to test takers' unique cultural and linguistic heritages, but we feature principled assessment design approaches as potential paths forward in this area. Following this, and referenced throughout the section, we address the tension between score comparability and the need for flexibility in test design to foster more equitable assessments of individuals from different cultural and linguistic groups. We end with a proposed IUA and validity argument that may facilitate the design and validation of more culturally responsive assessments. We suggest testing practices for the design and use of assessments that foreground equity and fairness considerations. Addressing equity and social justice issues should lead to improved validity of score inferences and uses for

all forms of testing, including credentialing and licensure testing as well as selection tests, even though there will be several design challenges to overcome.

Testing as a Situated Cultural Practice

As a culturally situated and socially constructed process for understanding what test takers know and can do, testing may not adequately represent the construct as manifested in various cultural, linguistic, and social contexts. In the testing situation, the targeted construct is not necessarily assessed in the same way; rather, it reflects each individual's cognitive skills and knowledge as shaped through social and cultural experiences and ways of knowing (Hood, 1998; Mislevy, 2018; Nortvedt et al., 2020). Performances on tests “are a function not only of the items, tasks, or stimulus conditions but of the *persons* responding and the *context* of measurement” (Messick, 1989, p. 14). The knowledge and processes evoked by test takers when performing a task depend on the test taker's cultural, linguistic, and experiential histories as well as the testing situation (Mislevy, 2018). To date, testing generally has overlooked the critical role of situated knowledge, disregarding how cultural and linguistic experiences and practices shape what individuals know and can do. Assessment of test takers from different cultural backgrounds bears on validity and fairness issues in score interpretation and test use because of differential access to, and representation of, the targeted construct (Camilli, 2006; Solano-Flores, 2011).

Tests designed to reflect how individuals learn and express their understandings, and account for different knowledge systems, can provide more valid score meaning and use (Trumbull & Nelson-Barber, 2019). The required backing in support of score meaning and test use for individuals with diverse cultural, social, and linguistic backgrounds needs attention throughout the testing process, from defining the construct and laying out the test design specifications to examining the differential consequences of test use. As an example, Messick (1989) argued that standardization in test materials and administration could be a source of invalidity across different cultures because “individuals from different cultural backgrounds might display their skills or engage task processes more in line with the respondent's cultural context” (p. 57). We acknowledge the need for many licensure and certification tests to adhere to common industry standards (e.g., pilot certification tests) requiring specific ways of demonstrating high levels of knowledge and skills. However, most educational applications can embrace cultural and linguistic heritages and practices to improve the representations of how all students come to know what is being assessed.

The equity and fairness concerns raised by Solano-Flores (2011) bear on the validity of score meaning and test use. Tests are designed to meet particular social needs, including compliance with laws and regulations; they are written in the dialect used by test developers; and test content reflects the knowledge, skills, and communication styles valued by the dominant culture. There is an inherent assumption that test takers are familiar with task contexts, ways of representing problems, and expected ways of responding (Solano-Flores, 2011, p. 3), yet each test taker approaches test items

through their own lens, drawing on their social and cultural experiences and ways of knowing. Experiences of the test takers during an assessment, and consequently their performances, are affected by a complex interaction of test characteristics (e.g., academic content, test language, item type, scoring) and cultural, language, economic, and educational histories of the test takers (Abedi, 2006; Solano-Flores & Trumbull, 2003). Language is central to the design of more equitable assessments because it is associated with the way test takers process, interpret, and understand the language of the task (Solano-Flores & Trumbull, 2003). The need to code-switch—altering tone and language depending on the context—can have undesirable effects on performance (Gregory & Fergus, 2017). Rater language background can be a source of construct-irrelevant variance because raters may interpret test-taker performance through their own cultural and linguistic lenses (Kachchaf & Solano-Flores, 2012). These issues are grounded in the recognition that testing is a situated cultural and social practice, and ways of knowing by test takers are culturally bounded. Both have significant implications for how IUAs and validity arguments are framed and how validity evaluations are conducted.

Addressing these concerns requires thoughtful attention to validity, equity, and social justice issues throughout test design and use, as well as cognizance of how values are reflected in the design, development, implementation, and use of tests. The design of culturally responsive and sustainable assessments requires attention to cultural characteristics as well as cultural assets of individuals, affording the opportunity for individuals from different cultural and linguistic groups to demonstrate what they know and can do. An important goal of culturally responsive assessments is to minimize disparate testing outcomes by race, culture, and language dominance due to the effects of the tests and testing practices. Of course, addressing testing issues is only one step of many to better serve a multicultural society.

Situated Learning, Assessment, and Validity

Modern conceptions of learning expand the range and nature of the evidence that must be addressed in validity arguments, even if the core of validity remains largely unchanged (Mislevy, 2016). Wilson (2018) and Shepard et al. (2018), among others, argued that the validity of assessment results can be enhanced if they are interpreted through a lens that connects assessment and instruction to how students learn. Not surprisingly, Cronbach and Meehl (1955) foreshadowed this work with their conception of the nomological network. The concept of a network makes clear that tests operate within educational systems, and their use cannot be validated without understanding how students are expected to progress through the domain in addition to other aspects of students' cultural and experiential backgrounds and instructional systems.

Socially Situated Learning Theories and Assessment Practice

Sociocultural (e.g., Penuel & Shepard, 2016a; Shepard et al., 2018) and sociocognitive (e.g., Mislevy, 2016, 2018) learning theories are two of the most prominent frameworks researchers are using to reconsider the foundations of educational assessment.

As Mislevy (2016) noted, the “socio” describes “the knowledge and activity patterns that structure individuals’ interactions with the world and others” (p. 267). Mislevy (2016) described three key types of patterns—linguistic, cultural, and substantive—that people learn for work and school, such as concepts and practices. Shepard et al. (2018) effectively summarized sociocultural theory, particularly in how it goes beyond cognitive development:

Sociocultural theory offers a powerful, integrative account of how motivational aspects of learning—such as self-regulation, self-efficacy, sense of belonging, and identity—are completely entwined with cognitive development. Unlike the social psychological and developmental psychology literatures, where each of these variables and related interventions are studied separately, sociocultural theory helps us see how knowledge and self-beliefs are jointly developed in communities of practice, and correspondingly how sense of self and meaningful participation may be harmed in unsupportive learning environments where children may be positioned as unable or deficient learners. (pp. 23–24)

Sociocognitive models are somewhat more limited than sociocultural models in that the former attend more to linguistic, cultural, and substantive patterns, whereas the latter attempt to account for the unique cultural and historical identities individuals bring to learning and motivation. This difference allows sociocognitive models to be a useful foundation for creating and understanding learning progressions based on “common starting points rather than making students’ interests and community experiences part of how interventions are designed” (Shepard et al., 2018, p. 25). However, this requires validity evidence demonstrating the applicability of learning progressions across diverse groups.

The Role of Situated Learning Theory in Assessment Validation and Validity

Validity is judged in relation to the inferences drawn from assessment responses, but how students come to learn the content of the test can influence the inferences one can draw. Insights from situated learning theory should be incorporated into the IUA. It is relatively straightforward to see how claims related to performance, scoring, and generalization inferences would benefit from an examination of how students are expected to come to know the targeted constructs.

Mislevy (2016) describes three types of evidence useful for supporting claims: student performance features, characteristics of the assessment situation, and other information about the student in the assessment situation. Evidence about student performance generally forms the basis of most validity investigations. Considering the linguistic, cultural, and substantive patterns indicated by sociocognitive theory provides richer understandings of the assessment situation, such as how test takers might interact with complex tasks like performance-based assessments.

There has been little attention paid to evidence related to the interaction of the student in the assessment situation, although such information is critical for understanding equity and fairness in assessment, as well as understanding the different evidentiary bases of contextualized classroom assessments compared with more distal large-scale achievement tests (Mislevy, 2016). That said, adhering to a sociocultural rather than a sociocognitive model could very well challenge the validity of inferences from standardized tests that, by design, ignore the uniqueness of individual students. However, Shepard et al. (2020) offered a set of principles to guide the design, implementation, and use of classroom assessment and instructional systems using a sociocultural lens.

Research is needed to explore the interaction among sociocultural theory, large-scale assessment, and validation. Recent work on the validation of instructionally relevant assessments (Pellegrino et al., 2016) is a promising step in this regard. Evaluators must adopt a critical stance when evaluating claims that culturally responsive assessments are more equitable. A strong logical and empirical program of research is needed to address the validity and equity of the assessment of diverse student populations at all levels of an assessment system.

Comparable and Valid Score Meaning

Employing a sociocultural or sociocognitive perspective for test design raises significant challenges for validating large-scale assessments. Comparability of scores is a key use case of large-scale assessments; consequently, the validity of score inferences and uses for most large-scale tests requires evidence of comparability across relevant entities (e.g., students, schools, states). Measurement invariance, a critical assumption in large-scale testing, means the same construct is being measured across different groups (e.g., different cultural and linguistic groups), resulting in the stability of item parameters across these groups. If this assumption is untenable, the validity of score interpretations and uses may be compromised. Principled test design procedures are intended to ensure comparable score meaning for individuals and aggregated groups. Depending on the IUA, aspects of comparability that may need to be examined across different groups include the following:

- response processes elicited
- precision of scores
- accuracy of decisions
- internal structure of the test
- accuracy of equating and scaling
- generalizability of the scores
- relationship between test scores and other measures
- intended and unintended consequences of score interpretations and test uses

For examples, in this section we address how analyses of internal structure and response processes can be adjusted to evaluate the validity of test score interpretations and uses in a more culturally responsive testing context.

Accounting for the heterogeneity of cultural and linguistic groups poses a challenge to these types of validity investigations. Future practices would benefit by reconceptualizing traditionally held beliefs about which individuals constitute a group. At a minimum, the variability of each group must be considered in the interpretation and use of the results.

The validity of score interpretations and uses, however, does not necessarily require all empirical relationships to be consistent across groups because of several factors, including different score ranges across different groups (Linn, 1983). Restriction of range can account for lower reliability estimates and poorer predictive validity evidence for some groups. Distributions of some test-taker groups may be centered in a different region on the score scale as compared with the general test population, and any lack of measurement invariance therefore may be due to nonoverlapping proficiency distributions rather than group status. Examining measurement invariance using a distribution of test takers from the general population that is similarly located on the score scale as the distribution of the group of interest provides more accurate and meaningful results (Buzick & Stone, 2011; Lane & Leventhal, 2015; Quesen & Lane, 2019; Sireci, 2009).

Internal Structure Evidence

Invariance of the internal structure of the test across different groups, and similar performance on items across different groups of similar competency level, support the validity of score meaning and use. DIF can be conceptualized as multidimensionality in test data. DIF could be a shift in the distribution of ability along a secondary construct that influences the probability of a correct response (Camilli, 1992). For example, one group may be more able on a secondary construct, such as English reading skills, when taking a mathematics test. Zwick (this volume) describes a variety of approaches for examining DIF and item bias that can shed light on differential validity of score interpretations and uses. Another approach to evaluate whether an item behaves similarly across different groups of test takers uses person fit indices based on item response theory (see, for example, Engelhard et al., 2014).

Treating a cultural or linguistic group as homogeneous in measurement invariance investigations may inaccurately portray the degree of score comparability (Ercikan et al., 2014; Oliveri et al., 2014; Solano-Flores & Li, 2009). Using Programme for International Student Assessment (PISA) data for Canada, for example, Ercikan et al. (2014) detected a lower percentage of DIF when conducting separate analyses for students who speak French at home and those who do not as compared to when conducting the analyses for the combined group. Such results underscore the need to consider the heterogeneity within each of the linguistic groups. A review of the DIF sources revealed that the linguistic load of the item, vocabulary, and complexity of sentence structure may be factors that disadvantage linguistic-minority students. While there has been more research on the heterogeneity of linguistic and special education student groups, the within-group diversity for racial, ethnic, and cultural groups must be attended to as well.

Response Process Evidence

The processes that test takers use to interpret, understand, and respond to test items are mediated by cultural, linguistic, and experiential factors. The interaction among test-taker characteristics (e.g., culture, language, learning style) and item features (e.g., problem context, graphics, language, response mode) deserves attention in test design and validation. Because different groups of test takers may employ different cognitive processes and strategies when responding to tasks, this differential use of response processes and strategies should be considered in the design of assessments. Similarly, response processes should be evaluated for different groups to inform the interpretation of across-group differences in score meaning (Lane, 2017). Cultural, linguistic, and experiential factors may influence performance in ways not anticipated by the test designer, which can impact the validity and fairness of score meaning and use for different groups. Therefore, cultural and linguistic factors that shape an individual's understanding of a construct must be considered in the design phase. Response process studies may challenge score meaning or the conceptualization of the construct.

Using think-aloud methods, Winter et al. (2006) showed that test-taker characteristics and item features interact to affect access to mathematics problem-solving tasks for English learners. Test takers were more likely to use a successful problem-solving process when items were varied to support better access to, and understanding of, the item. When items allowed for multiple modes of representation, students were more able to demonstrate their knowledge and skills. Using a response process analysis, Proser and Solano-Flores (2011) illustrated that English learners and non-English learners comprehended text differently for science and nonscience content terms and in the way they approached and understood more linguistically challenging sentences. Conducting think-aloud sessions, Ercikan et al. (2010) found that differences in item language in multilingual assessments explained differences in relative item difficulty for different groups. Kopriva and Wright (2017) proposed a conceptual framework for evaluating response processes in the assessment of nonnative speakers based on the premise that test design and score meaning requires an evaluation of the interaction between construct-irrelevant factors that test takers experience when responding to items (e.g., production skills) and conditions that are barriers in accessing the intended content and processes, including linguistic, cultural, and multisemiotic features (communication mode). These researchers proposed a system allowing test takers to use multiple modalities (e.g., nontext), enabling them to better access the assessment and, in turn, communicate their knowledge and skills. By better understanding how test takers use response processes when solving tasks (particularly complex problem-solving tasks) and how these processes may vary across different groups, test developers can improve test design and ensure valid and fair score interpretations and uses for different groups.

Comparability and Culturally Responsive and Sustainable Assessment

A major challenge when implementing culturally responsive and sustainable assessments is achieving the right balance between sufficient flexibility in addressing the needs of diverse cultures and sufficient standardization to meet comparability claims. Comparability among scores is required to make inferences about the relative performance of students and the effectiveness of programs and to monitor the progress of individuals or cohorts. Comparability generally requires some amount of standardization in assessment procedures and design, such as in the content and skills assessed, administration conditions, and scoring rules. Such standardization should support comparable and valid inferences within and across schools and over time (DePascale & Gong, 2020; Keng & Marion, 2020; Lane & DePascale, 2016).

However, flexibility in the design, implementation, and scoring of assessments arguably will better match how students have come to know the content and skills being assessed and, therefore, allow more valid score inferences and equitable assessments. An IUA delineating the intended inferences and assumptions also must describe the required comparability across tasks, across subpopulations, in scoring, and in administration conditions for the stated score interpretations and uses. To obtain a desired level of comparability, assessment features that can be altered must be specified and have accompanying guidance for altering those features (Lane & DePascale, 2016). The assessment designer must pay close attention to features that would allow students to demonstrate different ways of knowing and doing based on their culture, language, disability, and other relevant personal, social, and experiential characteristics. This discussion has focused on large-scale (e.g., state) assessments. Classroom assessments, as well as other tools intended to improve teaching and learning, should adopt more tailored approaches to address validity, equity, and social justice concerns (Shepard et al., 2018).

The Challenges to Validation

Advances in situated learning theory are helping to usher in a new era of educational assessment that holds great promise for improving the validity of score inferences. Similarly, designing assessments to reflect varying cultural and linguistic heritages promotes more valid interpretations and uses and supports more equitable education systems. These advances have considerable implications for validity, but they also complicate the drawing of defensible inferences from students' test scores. Mislevy (2016, p. 266) outlined eight challenges to validity posed by adopting a sociocognitive approach to assessment design and development:

- What do we want to assess?
- What kinds of performances do we need to observe, and in what kinds of situations?
- How should we think about constructs?
- How do we assess higher order skills?

- What is the role of measurement models?
- How do we “score” complex, interactive, performances at scale?
- How do we assess interactional skills?
- How do we take advantage of complex performance tasks?

Mislevy (2016, 2018) reflected on how operating from a sociocognitive perspective highlights the resources individuals have developed to engage in activities prescribed by the linguistic, cultural, and substantive patterns as practices valued by the community. Recognizing the heterogeneity of each cultural and linguistic group, he acknowledged that different linguistic, cultural, and substantive patterns may emerge within a particular cultural and linguistic group. Additionally, he and others (e.g., Ercikan & Pellegrino, 2017; Leighton, 2017) raised concerns that current testing is almost exclusively focused on the final product, whereas sociocognitive and sociocultural theories stress the importance of understanding the processes students pursue to shape their understandings on their way to a final product. Again, it is critical to understand and support these learning processes for almost all K–12 educational assessments.

Mislevy (2016) suggested that, from a sociocognitive perspective, constructs are ways to predict patterns of acting in common situations (or at least somewhat common) based on the resources individuals have developed. In this sense, a construct is as an “aggregation over clusters of individuals’ unique constellation of resources” (Mislevy, 2016, p. 274). Further, Mislevy (2016) noted, a sociocognitive perspective emphasizes that “there is no simple one-to-one correspondence between constructs, assessments, and measures” (p. 275).

Working from a sociocultural perspective, Solano-Flores (2019) provided a framework—more specifically, a “matrix of evidence for validity argumentation”—for evaluating the extent to which the practices and artifacts in a large-scale testing program address cultural diversity. His framework serves as “a tool for systematically gathering, contrasting, and integrating pieces of evidence that confirm or disconfirm the validity of interpretations and use of test scores for a culturally diverse population in a large-scale assessment program” (Solano-Flores, 2019, p. 7). It crosses 6 assessment process components and artifacts with 10 procedural assumptions. The framework serves as criteria for evaluating the rigor of the testing process. The process components include, for example, policies, uses, and consequences, whereas procedural assumptions include, for example, the specification of the population reflecting different cultural and linguistic groups; the inclusion, representation, and sampling of the cultural and linguistic groups; and the intersectionality of test data with other information about cultural and linguistic groups informing test use. Although Solano-Flores (2019) developed this conceptual framework in the context of large-scale educational assessments, much of it is applicable to other types of assessment.

Burns et al. (2019) specified general criteria used by university faculty in the United Kingdom for the design and use of culturally responsive assessments, which are applicable to other classroom assessment contexts as well. To ensure that these assessments yield more valid score interpretations and uses, criteria for the design

process include the consideration of cultural differences in how students interact with the teacher (as an expert authority or as an equal), engage with learning, approach knowledge, collaborate, and experience failure, while also attending to their previous real-life experiences.

Using Situated Learning Theory to Enhance Validity and Validation

Beyond improving how transfer and generalizability are understood, modern theories of situated learning will improve how we design assessments and investigate validity. We have witnessed since the beginning of the 21st century, considerable theoretical and empirical advances in principled assessment design for improving the validity of inferences from assessment results (Mislevy & Haertel, 2006; Mislevy et al., 2003; NRC, 2001; see also Huff et al., this volume). Principled assessment design has been employed purposefully to connect advances in learning theory with assessment design so that validity is incorporated from the design phase and is not an afterthought. We will consider two instantiations of principled assessment: evidence-centered design (ECD) and universal design for learning (UDL). The use of principled approaches allows for intentional and systematic design and validation of assessments that is responsive to different cultural, linguistic, and educational groups.

ECD was proposed as a framework for test design and interpretation that would provide better validity evidence to support inferences from test scores (Mislevy et al., 2003). ECD has three models that are sources of validity evidence: a student model, an evidence model, and a task model. The student model includes the domain modeling of the construct that is the intended focus of assessment. (Although Mislevy refers to *the student model*, this can be considered *the test-taker model* in other testing contexts, including licensure and certification testing.) The task model describes task characteristics (e.g., work products, demonstrations) and the environments in which students produce the desired evidence. Variable features in the task model can influence task difficulty and cognitive complexity. The evidence model, which links the task and student models, describes the evidence necessary to evaluate the student model and how that evidence should be considered when determining whether students mastered the intended knowledge and skills. The evidence model directly connects to the assessment argument by identifying the critical features of performance data scored through rubrics or other means.

UDL is a powerful framework for addressing issues of access and representation in learning and assessment (Dolan & Rose, 2000). Originally conceived to support access to learning and assessment for students with disabilities, UDL principles include the need for multiple means of representation, action and expression, and engagement (Thompson et al., 2002). UDL is now used to address how individuals from different cultural and educational backgrounds can demonstrate their understanding. The application of UDL principles throughout every phase of test design and implementation affords accessible assessments by minimizing

construct-irrelevant barriers that may negatively affect test performance. Moreover, UDL principles can attend to cultural assets to further minimize linguistic and cultural barriers to test performance. The application of universal design principles may differ for large-scale assessments as compared with classroom assessments, where task features can be modified more easily to account for differences in student experience and background. By including constructed-response items, however, large-scale assessments also can be responsive to different cultural and social experiences by permitting multiple means of demonstrating what individuals know and can do. Universal design principles are relevant in the design of tests for other purposes as well, such as licensure and certification.

Mislevy (2018) proposed applying the UDL principle of multiple means of action and expression by modifying scoring rubrics in light of cultural and linguistic considerations and, further, by conditioning differential familiarity into psychometric models using Bayesian networks. The UDL principles of multiple modes of representation and multiple means of engagement inform test design by conditioning task situations on information about test-taker background. Task models allow for variants of content that are more familiar, and of potential interest, to test takers, which provides a “conditional approach to inference” (Mislevy, 2018, p. 241). The use of task models allows for features of task and evaluation procedures to “be varied with respect to construct-irrelevant features across which individuals differ, in ways that leverage the resources each examinee has developed” (Mislevy, 2018, p. 241). Although test takers may be assessed with different variations of a task, each task variant aligns with the construct. Mislevy (2018) advocated enhanced construct representation through a greater reliance on the “theory of the construct,” even at the cost of reduced comparability (p. 242). This approach can enhance the validity of score meaning and use for test takers from different cultural, educational, and social backgrounds. We acknowledge the challenges that such an orientation poses to traditional test design and analysis, but if the validity of score inferences and test use is to be improved, taking us one step closer to social justice in testing, then student diversity must be addressed seriously throughout the testing process.

To be sure, approaches like those described above would be difficult to enact with some tests such as certification and licensure tests, compared with assessments used to inform instruction and student learning. However, even the design, scoring, and analyses of the former should be guided by equity and fairness considerations. The individualization and contextualization required for certification and licensure tests, however, likely will be considerably less than for assessments used to support and evaluate learning in schools.

To achieve fairness and equity in testing, collaboration is required among multiple parties, such as assessment specialists, psychometricians, learning theorists, content experts, and sociocultural experts. Indeed, the design of culturally responsive assessments requires this collective knowledge. The assessment-design advice from Qualls (1998) still resonates in the third decade of the 21st century:

What is perhaps most obvious with regard to this discussion of culturally responsive assessment is the need for collaboration across all stages of development. One must realize, however, that the person most knowledgeable in subject-matter content is not necessarily those most knowledgeable about relevant contextual cultural influences. (p. 297)

IUA and Validity Argument for Culturally Responsive Assessments

IUAs that reveal differing, and sometimes competing, values can enhance score meaning and use for individuals from culturally and linguistically diverse groups. IUAs that account for the different ways of knowing and doing that are relevant to the construct, as well as potential construct-irrelevant barriers for different cultural and linguistic groups, position the design and validation of assessments to be more culturally responsive. The validity argument used to outline the evidence required to support or refute warrants and assumptions should provide backing for claims that the tasks are relevant and meaningful to different cultural and linguistic groups. Additionally, principled assessment design approaches tuned to account for the unique linguistic and cultural backgrounds of test takers can support advances in equitable testing practices.

Table 4.2 provides an initial example of a general IUA and validity argument, focusing on those aspects that are most relevant to the design and validation of culturally responsive assessments. We are aware that the educational measurement field is rapidly advancing its understanding of how to infuse social justice principles throughout the assessment design and validation cycle, and there remains much work to be done. The parts of Table 4.2 that will be deemed relevant for a particular assessment will depend on the specific interpretation and use argument. Our intention here is to foster discussions of principled practices for the design and validation of culturally responsive assessments. The inclusion of stakeholders representing relevant groups will enrich such conversations.

For example, we propose that the performance inference must attend to potential differences in how the construct is defined, depending on different values and perspectives. We also address how potential differences in how individuals from various cultural and linguistic backgrounds internalize task features and contexts may require flexibility in assessment design. The specific use case will determine the extent to which standardization of tasks, scoring, and administration can be relaxed. Evidence to support the scoring inference includes response process evidence for raters. Particularly informative in this regard are evaluations of how raters internalize and apply scoring criteria, which can uncover potential biases, especially those that are subtle and arise from different ways of demonstrating competence in the domain. The generalization inference is supported by studies showing that inferences from the score to the broader domain are warranted for different groups. This includes evaluation of the equating error for different groups (when sample sizes permit).

Table 4.2 Interpretation and Use Argument and Validity Argument Components for Culturally Responsive Assessments

Interpretation and Use Argument		Validity Argument	
Inference	Warrant	Assumptions	Backing or Evidence
Performance: Infer from tasks to performance	Tasks elicit intended KSAs for all test takers	• Defined construct is relevant for different cultural groups • KSAs selected to be assessed are culturally relevant to as many groups as possible • Tasks and rubrics reflect KSAs • Individuals from different cultural groups were not prevented from accessing the test content because of construct-irrelevant features	• Documentation indicating how sociocultural and sociocognitive theories of learning are used for defining the construct • Construct-definition documents, including specifications of how competing values were addressed in defining the construct • Relevant groups involved in defining construct • Documentation indicating how, and by whom, KSAs were identified for inclusion in the blueprint • Documentation of how principled approaches were used in test and task design, including UDL principles • Task and rubric design and development procedures and diverse-expert review • Test-taker response process evidence, such as think-aloud studies for different cultural and linguistic groups
Scoring: Infer from performance to the score	Scoring rule that assigns a score to test performance reflects KSAs for all test takers	• Scoring procedures are appropriate for all test takers and lead to intended score meanings • Rubrics account for different ways of knowing and understanding • Scoring rules and rubrics are applied accurately and consistently for all test takers • Administration procedures are appropriate for all test takers	• Documentation of the design of scoring procedures that account for different ways of knowing and demonstrating what test takers know • Diverse-expert review of rubrics • Studies of comparability of score meanings • Studies examining the appropriateness of flexibility/variations in design • Rater-response process evidence, such as think-aloud studies, to evaluate raters' internalization and application of scoring criteria in ways that reflect equity goals

(continued)

Table 4.2 (continued)

Interpretation and Use Argument		Validity Argument	
Inference	Warrant	Assumptions	Backing or Evidence
			• Rater-consistency and -accuracy monitoring for different cultural and linguistic groups • Evidence of consistency and accuracy of automated scoring algorithms for different cultural and linguistic groups • Evidence of the appropriateness of administration procedures
Generalization: Infer from score to the broader construct or domain	Allows for dependable inferences from performance to score to the expected performance in the broader domain for all test takers	• Test provides a sufficient representation of the KSA in the broader domain • Scaling and equating procedures are appropriate for different cultural and linguistic groups	• Generalizability studies for tasks, occasions, and other relevant facets for different cultural and linguistic groups • Studies of the internal structure for different cultural and linguistic groups (e.g., multi-group factor analyses, person-fit analyses) • Equating error is minimized and similar across cultural and linguistic groups
Extrapolation: Infer from test performance to expected performance in the real-world contexts	There is a relationship between test performance and using KSAs in real-world contexts for all test takers	• Task contexts are relevant for different cultural and linguistic groups • Relationships between test scores and real-world criteria are similar across cultural and linguistic groups	• Expert review of the alignment of KSAs tested to the KSAs used in diverse real-world contexts • Expected relationships between test scores and other criterion measures are similar across different cultural and linguistic groups
Decision/use inference and consequences	Decision rule and uses of test are appropriate for all test takers	• Assignment to performance levels is appropriate across cultural and linguistic groups • Decisions and uses of test results are appropriate and fair across cultural and linguistic groups • Use of tests leads to equitable consequences	• Standard-setting method and procedures are appropriate and address possible adverse impact • Decision consistency and accuracy are similar across cultural and linguistic groups • Evidence that there is a shared understanding by different stakeholder groups of intended positive consequences and potential unintended negative consequences (with any differences documented) • Decisions and uses do not result in negative unintended consequences for individuals from different cultural and linguistic groups

Note. KSA = knowledge, skills, and other attributes; UDL = universal design for learning.

The extrapolation inference calls for evidence that expected relationships between test scores and criterion measures hold for different groups. These are nuanced studies that should consider the reciprocal relationships between academic concepts and concepts developed within one's own culture and community. Consequences are at the core of the decision/use inference; however, unfortunately, there are few studies that have evaluated the consequences of testing programs. When setting performance standards, potential adverse impact should be investigated and minimized, and perceived positive consequences and potential unintended negative consequences held by different stakeholder groups should be documented and evaluated.

Standardization Versus Flexibility

The IUA that lays out the intended inferences and assumptions will shed light on the degree of comparability needed across tasks and student populations and in scoring and administration conditions. The required level of comparability will depend on the purpose and use of the assessment results. For scores to be comparable within a testing occasion and over time requires a level of standardization in assessment design, including the content and skills assessed, administration conditions, and scoring rules and procedures (DePascale & Gong, 2020; Keng & Marion, 2020; Lane & DePascale, 2016; for a discussion on administration and scoring, see Shermis et al., this volume). Departures from standardization in design features, such as task modifications, on the one hand may pose threats to score comparability but, on the other hand, may provide for more equitable assessments. Flexibility in classroom assessments may result in a better alignment with the goals of instruction and more fair and equitable assessments. Validity researchers and other stakeholders must adopt a critical stance when evaluating the claims that culturally responsive assessments are more equitable. A strong logical and empirical program of research is needed to address the validity and equity of the assessment of diverse populations.

CURRENT APPLICATIONS OF VALIDITY THEORY AND PRACTICE

We discuss in this section the application of validity theory to testing, including K–12 assessment systems and technology innovations in assessment design.

K–12 Assessments and Assessment Systems

There are many examples in the K–12 context showing how validity theory and practice are applied when evaluating the quality and usefulness of inferences from test scores. We highlight two applications: interim assessments and balanced assessment systems.

Interim Assessments

Interim assessments have proliferated since the passage of No Child Left Behind (NRC, 2010), to the point where most school districts in the United States use a commercial or state-provided interim assessment. This led Perie et al. (2009) to delineate the variety of terms and uses associated with these assessments, which are situated somewhere between classroom- and state-level assessments. Perie et al. (2009) coined the term *interim* to distinguish between (a) formative assessments having research-based uses to inform instruction and improve student learning (e.g., Black & Wiliam, 1998) and (b) other assessments, which are administered more formally and controlled outside the classroom and lack a research base to support their use.

Perie et al. (2009) documented the many purposes interim assessments are purportedly designed to serve, which they categorized as instructional, evaluative, and predictive (see Table 4.3).

Validation occurs in the context of specific uses to serve particular purposes. There is little independent empirical evidence in support of the quality of interim assessments (e.g., Dadey & Diggs, 2019) or in support of any of the uses for interim assessments summarized in Table 4.3 (e.g., Goertz et al., 2009; Konstantopoulos et al., 2016). Validating score inferences and uses for these purposes would be a considerable undertaking. Nonetheless, if providers claim their assessments support these uses—and do so efficiently and effectively—these providers are obligated to provide the corresponding validity evidence.

IUAs for Interim Assessments

The choice of an interim assessment, whether by state, district, or school leaders, should be driven by the intended purpose of the assessment (Perie et al., 2009). An interim assessment chosen for instructional purposes should be linked to specific instructional

Table 4.3 Interim Assessment Uses

Category of Use	Most Common Stated Uses
Instructional	• Modify or improve instruction • Identify students needing additional support • Identify content to reteach • Differentiate instruction • Document individual student growth within a school year
Evaluative	• Monitor program initiatives • Monitor performance for accountability • Evaluate curricular programs • Monitor performance trends of identified student groups • Improve overall school performance
Predictive	• Forecast performance on the end-of-year summative assessment

units, thereby providing teachers with meaningful results that inform instruction and student learning. Such assessments likely are more aligned with local instruction and assessments than other types of interim assessments. An interim assessment chosen for evaluative purposes should provide information for modifying and improving curriculum and, consequently, instruction. An interim assessment chosen for predictive purposes should provide information regarding how well the student is likely to perform on the end-of-year assessment. Depending on their purpose, interim assessments should have different designs and require different validity evidence to support score meaning and use. We focus here on the purported instructional uses of interim assessment.

Pellegrino et al. (2016) provided a helpful framework for evaluating the validity of test score inferences from assessments intended to support instruction and student learning. They identified three types of validity—cognitive, instructional, and inferential—that should be used to structure validity arguments and evaluations of such assessments. Although we do not agree with their terminology—their three categories arguably are types of evidence, not types of validity—we believe their framework can be applied to the validation of interim assessments intended to improve teaching and learning. Cognitive evidence refers to how an assessment elicits evidence of test takers' KSAs and ties to an underlying research-based model of learning. Instructional evidence reveals the extent to which the assessment is aligned with the enacted curriculum and students' opportunities to learn the targeted knowledge and skills. This facet also includes evidence regarding how well assessment results provide insights into student understanding, enabling educators to adjust their teaching practices accordingly. This important use clearly overlaps with the consequential evidence that a more typical IUA speaks to. As such, instructional evidence should be collected to determine whether, and how, the assessment results lead to missed instructional opportunities or more egregious unintended negative consequences. Finally, inferential evidence is important for evaluating the use of assessment score inferences for instructional purposes. As Pellegrino et al. (2016) stated, this line of evidence focuses on “the extent to which an assessment reliably and accurately yields model-based information about student performance, especially for diagnostic purposes” (p. 62).

These three forms of evidence overlap with the types of evidence called for in the 2014 *Standards*. Further, they should be integrated into an IUA to support a pragmatic approach for evaluating the validity of assessment results and uses that are intended to support teaching and learning. Most interim assessment providers make strong claims about the utility of their assessments to improve student learning by better directing teacher actions. Such claims require strong evidence, along the lines Pellegrino et al. (2016) argued for. Unfortunately, there is essentially no independent evidence to support the strong claims made by the providers of interim assessments (e.g., Dadey & Diggs, 2019).

We have stated throughout this chapter that validity is evaluated in light of specific intended interpretations and uses. Ing et al. (2021) recently emphasized and extended this point. They provided compelling evidence that one cannot simply assert that validity evidence transfers from one purpose to another. Rather, validation must attend to specific contexts, which is important advice for interim assessment providers to bear in mind. Just because a set of interpretations and uses are justified in a wealthy suburban district, for example, does not mean these interpretations and uses are therefore justified in a different context.

Balanced Assessment Systems

There has been a long-standing belief that a balanced and integrated system of assessment, curriculum, and instruction would inform instruction and enhance student learning (Pellegrino et al., 1999; Shepard, 2000). Shepard (2000) discussed the need to reconceptualize classroom assessment practices so that they support student learning. This includes changing the content and form of classroom assessments to better reflect critical thinking and problem-solving skills and changing what information is collected to better inform the teaching and learning process. Coherence across classroom assessments and large-scale assessments requires a reconceptualization of classroom assessments, as well as a restructuring of state assessment and accountability programs. A balanced assessment system is achieved when the assessments in the system are coherently linked through a clear specification of the learning targets, comprehensively support multiple purposes and uses, and continuously monitor student learning over time (NRC, 2001). These properties of coherence, comprehensiveness, and continuity inform the design, use, and validation of high-quality balanced assessment systems (NRC, 2001).

Research-based theories of situated learning must form the basis of large-scale assessments and classroom assessments to ensure coherence across them and to improve the instruction and learning process (NRC, 2001; Shepard et al., 2018; Wilson, 2018). This requires tailoring the shared theories of situated learning to the assessment purpose and use, where the corresponding model of learning could be broader for large-scale assessments that monitor trends and are used for accountability purposes and finer grained for assessments used to inform instruction and improve student learning. Theories of learning situated in a social context support teaching and assessment practices that recognize different ways of knowing and doing and, consequently, can promote equity and social justice (Penuel & Shepard, 2016b). Sociocultural learning theories must form the basis for these balanced assessment systems, requiring that both assessments and instruction are responsive to different ways of knowing, shaped by the social and cultural experiences of students. Flexibility in the design of formative assessments allows for the experiences and backgrounds of students to be considered. However, evidence should be collected to support, or potentially refute, that such flexibility in assessment design leads to more valid score meanings and uses for the intended subpopulation.

A Tightly or Loosely Coupled System

Marion (2018a) suggested that designing a comprehensive assessment system requires a “loose coupling” (p. 45) between large-scale and local assessments, relaxing the criteria of a “tightly coupled system” advocated in *Knowing What Students Know* (NRC, 2001). A tightly coupled system may threaten the validity of score meaning and use for assessments at one or more of the levels of the system. If classroom assessments mirror large-scale assessments—which measure a restricted depth of KSAs—the former will therefore provide limited useful information for improving instruction and learning. Marion (2018a) provided examples of states that have loosely coupled systems where information from one level (e.g., classroom) does not necessarily support the purposes at another level (e.g., accountability). This allows for the assessment of the content more deeply at appropriate levels of the system to provide more meaningful information to assessment users. An IUA that specifies the claims for each level of the assessment system allows score interpretations and uses to be tailored to each level.

IUAs for the Assessment System

A multilevel IUA is required for evaluating the validity of an assessment system. Each assessment (or at least the major assessments) in the system must be evaluated in view of the assessment’s intended interpretations and uses, as articulated in the IUA. The IUAs for the assessments in the system must address the degree to which each assessment contributes to supporting the quality of the system as a whole. A prominent challenge for large-scale summative assessments is to produce score information that is explicitly tied to the specific content and skills students are expected to learn (i.e., content referenced growth). *Knowing What Students Know* (NRC, 2001) addressed this challenge over 2 decades ago, but, unfortunately, little progress has been made in this regard (Marion et al., 2019).

As we have noted, the NRC proposed three criteria—coherence, comprehensiveness, and continuity—for evaluating the balance of an assessment system (NRC, 2001). These criteria have not been operationalized so that the field has sufficient guidance for judging, say, how much coherence is required for concluding that an assessment system is balanced. Nevertheless, an assessment system’s IUA must clearly delineate the goals and purposes, intended score interpretations and uses, and intended consequences and potential undesirable consequences of each assessment in the system and the system as a whole. The system-level IUA should articulate claims regarding coherence, comprehensiveness, and continuity in terms of overall system goals. Further, because assessment systems are designed to support multiple users and stakeholders in the educational system, the utility and consequences of assessment scores must be central to any validity evaluation of the assessment system.

There is no question that creating and evaluating a multilevel IUA will be challenging for validity evaluators and assessment system stakeholders. Like other complex evaluations, validity evaluators should begin with the system’s highest-priority goals. In the case of assessment systems, these often will address considerations regarding utility and consequences, while the evaluation of score interpretations generally will be based on evaluating arguments associated with individual assessments.

Technology Use in Testing Programs

Many testing programs make use of technology, such as technology-enhanced items, automated scoring, and computer-based simulations, which provide opportunities for enhancing the validity of score meaning and test use (see Bennett et al., this volume). The rapid development of technology provides both opportunities and challenges to the design, development, implementation, and use of assessments. For example, a key advantage of computer adaptive testing is its capability to select tasks appropriate to the test taker's competency level. However, if the content specification targets used to guide this selection are too coarse, it threatens the representation of the assessed construct for the test taker and, therefore, compromises the validity of score interpretations and uses. Evidence to support the construct representation of a computer adaptive test requires an evaluation of the alignment of the test that test takers actually receive with a more refined, meaningful level of test specifications.

The use of technology in testing programs varies, from the design and development of technology-enhanced items (e.g., drag and drop, audio and video clips) to simulation tasks and game-based assessments. The latter require innovative assessment designs that present dynamic, engaging interactive tasks assessing complex knowledge and skills. In the assessment of language learners and students with disabilities, technology can promote access to test content by providing accommodations that enhance the validity and fairness of score interpretations and uses (E. Stone et al., 2016), but it also can introduce a new source of construct-irrelevant variance.

Recent technological advances, such as educational data mining and machine learning, have potential to assess difficult-to-measure competencies, competencies based on cognitive models of achievement and learning, and competencies that require complex problem-solving, strategies, and reasoning. An overarching goal in the use of these technological advances in test design is to improve construct representation while minimizing construct-irrelevant variance. Educational data mining and machine learning can capture salient features of the processes and actions used by tests takers. They are used in the design of simulation-based assessments, game-based assessments, and collaborative problem-solving assessments with the goal of obtaining rich data on individuals' actions and processes in real time for primarily formative purposes. These advances, however, must be critically examined for potential threats to validity.

Technology Use in Capturing Response Processes

Researchers have leveraged technology in designing assessments that better measure response processes critical to claims regarding complex problem-solving. Technological advances allow for capturing evidence of the processes underlying performance in real time and for assessing a wider range of test-taker behaviors, thereby improving construct representation. Response process data are essential to the design, development, and interpretation of assessments measuring complex cognitive constructs (Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017). Oranje et al. (2017) discussed three types of data—logs, response time, and eye tracking—that are indicators of underlying processes

and can therefore shed light on score meaning. In game-based learning environments, for example, data logging and educational data mining can capture behavioral events that link game play to the individual's knowledge, cognitive skills, and strategy use. Statistical methods such as cluster analysis can then be used to analyze and summarize the logged behaviors by extracting clusters of behaviors representing different cognitive patterns (e.g., different strategy use). These data types can shorten the inferential distance between claims and test performance by testing proposed inferences pertaining to the intended construct. These data types also can improve test design and user interfaces, provide a context for the inference about the intended construct, serve as indicators of engagement, and remove noise from the process data (Oranje et al., 2017).

Process data obtained throughout test design, development, and implementation can inform the validity of score interpretations and uses by identifying potential sources of construct-irrelevant variance and providing evidence to support (or refute) process models. Although technology-based assessments can produce a large amount of data, one must determine which data are relevant and how they are to be aggregated to justify score inferences and uses. Response process data are as important for disconfirming IUA assumptions as they are for corroborating them (Kane & Mislevy, 2017). Automated scoring, simulation-based assessments, game-based assessments, and collaborative problem-solving are examples of rapidly developing areas that are harnessing advances in technology and machine learning (Mislevy, 2018; Mislevy et al., 2012, 2014; von Davier et al., 2017).

Automated Scoring

As the field moves away from human raters to automated scoring, as in the evaluation of test-taker responses to virtual performance tasks and constructed-response tasks, a paramount validity consideration is the meaning of scores generated by scoring algorithms. Automated scoring involves traditional methods (e.g., regression) as well as more contemporary ones (e.g., natural language processing and machine learning) and has potential to improve the validity of score inferences and uses. Specifically, automated scoring applies the scoring rubric consistently, enables test designers to control the features considered in scoring, captures multiple aspects of performance, and allows for timely score reporting (Powers et al., 1998; Williamson et al., 2006). Further, its uniform application across individuals has the potential to promote score comparability. These benefits notwithstanding, automated scoring carries potential threats to validity. For example, the assessment is vulnerable to construct underrepresentation by only measuring KSAs amenable to machine scoring. Also, test length is a strong predictor of human essay scores; unless its weight is decreased in the automated scoring system, this will contribute to construct irrelevant variance.

Williamson et al. (2012) provided an argument-based validity framework for evaluating automated scoring, identifying the evidence needed to support scoring, generalization, extrapolation, and decision/use inferences. (This work extended the validity framework offered by Clauser et al., 2002.) A common source of validity evidence to support the scoring inference is the relationship between automated and human scores,

relative to the relationship between two human scores. Bejar (2012, 2017) argued that human scores require an appraisal of their own IUA before they can be used for the design and validation of automated scores. He referred to this as the first-order IUA for automated scoring. Such an IUA also requires evidence to support the human scores of responses used to train the scoring algorithm.

The sample of human scored test-taker responses for training and calibrating the scoring engine needs to be evaluated in terms of its representativeness of the potential types of responses and of the score scale when compared to expert human scorers (Bennett & Zhang, 2016). The sample of responses for training should be drawn to adequately represent different cultural, linguistic, and disability groups in the population of potential test takers to ensure that the algorithm does not favor one group's way of responding over others (Kolen, 2011). Kolen (2011) called for research to investigate the comparability of scores based on using test-taker responses from different cultural and linguistic groups to train the scoring algorithm. For example, one could examine the "difference in scores for African American students if the engine was trained only on African American students, only on White students, or on a combination of African American and White students" (M. J. Kolen, personal communication, June 17, 2019). Such a study would address the concern that assessments of writing proficiency privilege the American English dialect (Randall et al., 2022). Similar studies could be expanded to include other groups (e.g., Hispanics, English learners, Indigenous people, students with disabilities), moving one step closer to the design of more culturally responsive and sustainable assessments that consequently provide more valid score inferences and uses.

Other sources of validity evidence include the generalizability of automated scores to other tasks, settings, and contexts; the extent to which the expected relationships to external criteria hold; and the extrapolation from automated scores to performance on a real-life measure (Bennett & Zhang, 2016; Williamson et al., 2012). Clauser et al. (2002) discussed the degree of judgment inherent in the design of scoring algorithms, calling for investigations of the comparability and generalizability of scoring algorithms developed by equally qualified experts. Harik et al. (2013) studied different groups of content experts who developed rule-based scoring algorithms for virtual simulations. These researchers found a group effect, indicating that it matters what group of content experts is chosen for training the algorithm when using rule-based methods.

Culturally responsive design of scoring algorithms requires investigations of the comparability of score meaning across diverse groups. This includes appraisals of group differences regarding relationships with human scores and relationships with external measures, as well as evaluations of the extent to which generalizations and extrapolations are comparable across groups. In their discussion of threats to the validity of automated scoring, Clauser et al. (2002) indicated that features in the algorithms used to predict human scores can be sources of construct-irrelevant variance. For example, Bejar (2012) found that the extraction of features, and the synthesis of those features in the scoring engine for the GRE Analytic Writing task, were vulnerable to

construct-irrelevant test-taker strategies, such as adding words to increase the length of the response. Some test takers used strategies to maximize their scores in a construct-irrelevant manner, which posed a threat to score meaning and use.

Mao et al. (2018) provided backing for the decision/use inference of an automated scoring system for a formative assessment of science argumentation. Their results supported the assumption that feedback based on students' automated scores would prompt students to modify their responses and improve their scientific argumentation skills as they revised their responses.

Simulation-Based Assessments

Technological advances allow for creating and emulating complex real-life scenarios and provide opportunities to assess individuals' capabilities in such scenarios. Computer-based simulation tasks can entail the measurement of complex reasoning and inferential skills that are difficult to measure in traditional forms of assessment. Many of the advantages and challenges in the design, implementation, and use of simulation-based assessments are not unlike those of other performance-based assessments, such as threats to the generalization inference (Lane & Stone, 2006), but additional challenges and opportunities arise in the virtual environment. Credentialing and K–12 classroom-based assessment programs both have seen the implementation of virtual simulation tasks that vary in their level of complexity and interactivity. Simulation-based assessments, including game-based and collaborative problem-solving assessments, can capture individuals' interaction patterns to assess learning and provide scores for achievement and noncognitive domains alike.

IUAs for Simulation-Based Assessments

IUAs for simulation-based assessments depend on the particular use of the assessment, but there are common aspects across IUAs. Underlying the use of simulation tasks is the expectation of improved construct representation and improved extrapolation from the performance to the real-life context, given a key assumption: The KSAs required by the simulated performance reflect those used in the real-life context. Regardless of how well designed they are, simulation tasks are not identical to the real-life criterion insofar as the simulation task may model some features of the real task while excluding others. Capturing the full richness of complex performances may be challenging because of constraints in the virtual environment. Backing for the assumption that the simulation is sufficiently similar to actual performance would include evidence gathered from the task and scoring design procedures, such as cognitive models used in the design and their appraisal by content experts. The development of cognitive models that underlie test-taker performance is an iterative process, requiring ongoing validity research. Empirically and theoretically based cognitive models are not available or fully developed in most domains; thus, specification of the process features may pose challenges and affect the validity of score meaning and use.

Construct representation, as well as inferences from task to targeted performance, are supported through the use of ECD in that student models comprising the targeted

KSAs are linked to task models, and task models then are linked to evidence models that describe test-taker competency on the targeted KSAs. Construct representation rests on the alignment of (a) the elicited processes and behaviors and interaction patterns in the virtual assessment with (b) those processes, actions, and interaction patterns used in real-life contexts (Mislevy, 2018; von Davier et al., 2017). Explicit delineation of claims and evidence can reduce construct-irrelevant variance and help ensure sufficient and accurate representation of the intended KSAs and interaction patterns. Other sources of construct-irrelevant variance in the design phase requiring attention include irrelevant text and representations that lead to extraneous cognitive processing that may negatively affect performance (Quellmalz et al., 2013).

More equitable assessments will result from having different cultural norms and linguistic patterns represented in the design phase, along with attending to salient features of task design and construct representation (Mislevy, 2018). Including cultural and linguistic experts in the construct definition and design phase will help uncover, and then incorporate, construct-relevant cultural differences in interaction patterns, thus avoiding a source of construct-irrelevant variance and enhancing score meaning (Oliveri et al., 2019). Response process evidence from think-aloud sessions and cognitive labs can provide additional backing for construct representation, as well as for making inferences from the task to performance. Such evidence also can uncover potential sources of both construct-relevant and construct-irrelevant variance.

Clauser et al. (2016) examined the warrants and backing for scoring, generalization, extrapolation, and decision/use inferences for simulation-based assessments. A principled design approach can support the scoring inference because the identification of needed validity evidence is inherent in the design phase, allowing salient features of test takers' competencies to be captured (Clauser et al., 2016). Standardizing the administration of the task based on test-taker actions, and standardizing how scores are derived for test takers who have the same path when working on the task, can provide additional backing for the scoring inference (Clauser et al., 2016). However, such uniformity and standardization may restrict different pathways that reflect diverse backgrounds and, therefore, affect the validity of score meaning and use. Although not focused on linguistic and cultural factors, Zhai et al. (2021) recently advanced a framework for the validation of machine learning for the Next Generation Science Assessment initiative. Their framework addresses the complexity of applying and validating machine-learning processes in the context of multidimensional science assessments. They noted that machine learning may foster the assessment goals of Next Generation Science Standards (2013), but machine learning also may introduce construct-irrelevant factors threatening the validity of score interpretations and use. Addressing how machine learning algorithms interact with various cultural and linguistic backgrounds will undoubtedly pose additional design and validity challenges.

The aggregation of data to produce scores may be problematic in the absence of a cognitive model. While construct representation may be supported for a particular task in terms of assessing KSAs that are used in the criterion context, the generalization

inference may be jeopardized. As is the case with most performance assessments, a few simulation tasks typically are sampled from the domain, which limits the generalizability of the score inferences to this domain. In a formative assessment and learning environment, generalizability to the broader domain of tasks may not be a concern. Because test-taker performance and resulting scores are affected by the choices made in the design phase of simulations, appraising how inferences may vary with different design choices will provide additional backing for the scoring inference (Mislevy et al., 2016).

Validity evidence to support the extrapolation inference derives, in part, from the demonstrated alignment of simulation-task content with the knowledge and skills needed to perform the task in the criterion setting and the demonstrated relationship between assessment scores and performances in the criterion setting (Clauser et al., 2016). If test takers respond to the task differently in a testing context versus the criterion setting, this may contribute to construct-irrelevant variance (Clauser et al., 2016). Test takers may use different strategies and processes when responding to virtual tasks than when in the criterion setting. This may be less of a concern for well-designed, engaging game-based assessments used for formative purposes. Other potential sources of construct-irrelevant variance include the use of virtual agents, interface and representations that are possibly unfamiliar to the test taker, and the level of test-taker engagement and motivation. The interaction of these sources with the test takers' cultural, social, experiential, and personal background may further complicate validity. For example, while not assessment related, racial bias in facial recognition software (Harwell, 2019) arguably has implications for designing and implementing automated learning systems and the corresponding assessments. Undesirable consequences of the decision/use inference include narrowly focusing instruction on the assessment's scoring criteria instead of the KSAs in the broader domain (Clauser et al., 2016).

Collaborative Problem-Solving Assessments

Virtual collaborative problem-solving assessments, designed primarily for formative purposes, carry validity threats as well. For example, construct-irrelevant variance can result from variability in partner skill: A test taker may have insufficient opportunity to demonstrate maximum performance on a collaborative problem-solving task if their human collaborator has limited collaborative skills. The use of a virtual, rather than human, agent may provide the opportunity for test takers to better demonstrate their KSAs (Rosen, 2017; Scoular et al., 2017). Although virtual agents allow for uniformity in paths for those test takers who perform the same action, they may be a source of construct-irrelevant variance because they cannot adjust to unusual test-taker behaviors (Rosen, 2017). Further, the use of virtual agents may impede task engagement by test takers. When the intended population of test takers is diverse, test designers should consider how virtual agents are presented and the language they use (Oliveri et al., 2019). Think-aloud sessions can throw light on whether interaction patterns differ across culturally and linguistically diverse groups and, further, whether features of this interaction space differentially impede the performance of these groups.

Challenges to the scoring inference can arise when disentangling domain-relevant competencies (e.g., science inquiry skills) from collaborative skills, given the data. Supporting evidence is required to generalize test-taker performance across different groups of collaborators (e.g., categorized by skill level, cultural and racial background, gender), tasks (and task features, such as a virtual agent vs. a human collaborator), and occasions (Rosen, 2017). The assumption of an improved extrapolation inference for these assessments over more traditional formats requires support. The challenge of capturing the richness of the real-life interaction in a virtual environment affects construct representation. For example, in a study that modeled different patterns of interaction, where some patterns were more related to learning than others, Andrews et al. (2017) detected a “fake collaboration” pattern where both members of the dyad appeared to collaborate but maintained their own responses regardless of their prior agreement with the other’s response.

Technological advances in assessment design, implementation, and scoring have the potential to assess competencies that have been difficult to measure, assess complex problem-solving and inferential skills, and enhance validity and fairness of score interpretations and uses. However, a critical examination of potential threats to validity and fairness should be undertaken for technology-based assessments.

CONCLUDING COMMENTS

Validity is at the core of every aspect of testing, from conceptualizing and defining the targeted construct to evaluating the consequences of test use. The design of tests begins with an IUA that explicitly links the inferences from the tasks to the decisions and uses and then to the resulting consequences. The validity argument provides a comprehensive framework, including the rationales, assumptions, and backing or evidence, to support (or refute) the inferences and uses laid out in the IUA. IUAs and validity arguments must attend to fairness issues in the testing process.

Fairness issues are fundamental to the validity of score meaning and test use for individuals from diverse groups. We have argued that fairness issues are validity issues and, therefore, must be addressed in the IUA and validity argument. Testing programs that explicitly address how cultural and linguistic experiences and practices shape how individuals learn, know, and demonstrate their understanding can lead to more equitable assessments. Sociocultural theories of learning provide a foundation for the design of assessments that embody different ways of knowing. More valid score meaning and use can be achieved by leveraging technology and advanced psychometric methods in the design of tests to better capture processes and strategies that are construct relevant but may differ across population groups (e.g., Mislevy, 2018). A well-crafted IUA and validity argument must address scoring meaning and use for individuals from all targeted population groups.

We have proposed an approach for crafting IUAs and validity arguments in response to calls for more culturally responsive and sustainable assessments. While some aspects may be more relevant for educational tests than credentialing tests, credentialing tests

must also attend to fairness and equity issues. As an example, medical licensure tests for physicians typically assess communication and interpersonal skills. Assessment of these skills should reflect physicians' need to understand different cultural norms in communicative styles and expressive language. Such sections of licensure exams could assess how the test taker pays attention to and responds appropriately when interacting with individuals from different cultures.

We must accelerate efforts to improve the validity of score meaning and use for individuals with differing cultural, linguistics, experiential, and personal backgrounds. Although they pose challenges, advances in measurement models, sociocultural learning theory, and technology hold promise for the design and use of culturally responsive and sustaining assessments. We support the call for a justice-oriented framework for conceptualizing constructs and identifying construct-irrelevant sources of variance to better promote and assess diverse cultural and linguistic ways of knowing and doing (e.g., Gordon, 1995; Gordon Commission, 2013; Randall, 2021). Similarly, we must ensure that the design of the IUA and validity argument addresses any inherent biases so these arguments are designed to better ensure valid score inferences and test use for traditionally marginalized cultural and linguistic groups.

ACKNOWLEDGMENTS

We are grateful for the thoughtful reviews of the chapter provided by Joan Herman, Michael Kane, Paul Newton, and Lorrie Shepard. We are also in debt to Jennifer Randall and Guillermo (Willy) Solano-Flores for their insightful suggestions on the section "Assessment of Individuals With Varying Cultural and Linguistic Backgrounds." Our editors, Linda Cook and Mary Pitoniak, provided excellent support and advice throughout the process. We would also like to note that this chapter was completed in the beginning of 2022.

REFERENCES

- Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Lawrence Erlbaum Associates.
- Achieve. (2018, March). *Independent analysis of the alignment of the ACT to the Common Core State Standards*. Retrieved December 17, 2019, from <https://www.achieve.org/files/ACTReport.pdf>
- Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13, 280–297.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129. <https://doi.org/10.1093/applin/14.2.115>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*.

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing*. American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Educational Research Association and National Council on Measurements Used in Education. (1955). *Technical recommendations for achievement tests*. National Education Association.
- American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. *Psychological Bulletin*, 51(2, Part 2), 1–38.
- Andrews, J. J., Kerr, D., Mislevy, R. J., von Davier, A., Hao, J., & Liu, L. (2017). Modeling collaborative interaction patterns in a simulation-based task. *Journal of Educational Measurement*, 54, 54–69.
- Atkinson, D. (2012). Legal issues and considerations for standard setting in professional licensure and certification examinations. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 503–534). Routledge.
- Bachman, L. F. (2002). Alternative interpretations of alternative assessments: Some validity issues in education performance assessments. *Educational Measurement: Issues and Practice*, 21(3), 5–18. <https://doi.org/10.1111/j.1745-3992.2002.tb00095.x>
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, 2(1), 1–34. <https://doi.org/10.1207/s15434311laq02011>
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Barnes, M. (2016). The washback of the TOEFL iBT in Vietnam. *Australian Journal of Teacher Education*, 41(7), 158–2016.
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2–9.
- Bejar, I. I. (2017). Threats to score meaning in automated scoring. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 75–84). Routledge.
- Bennett, R. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment.

- Measurement: Interdisciplinary Research and Perspectives*, 8(2–3), 70–91.
<https://doi.org/10.1080/15366367.2010.508686>
- Bennett, R., Kane, M. K., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment* [Paper presentation]. Invitational Research Symposium on Through Course Summative Assessment, Atlanta, GA, United States.
- Bennett, R., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing* (pp. 142–173). Routledge.
- Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of large-scale educational assessments: Issues and recommendations*. National Academy of Education.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7–74.
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42(2), 231–268.
- Borsboom, D. Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brennan, R. L. (1992). *Elements of generalizability theory*. American College Testing Program.
- Burns, D., Brown, M., O’Hara, J., & McNamara, G. (2019). Progressing culturally responsive assessment for higher education institutions. In *Diversity and triumphs of navigating the terrain of academe: International perspectives* (pp. 63–85). Emerald Publishing Limited.
- Buzick, H., & Stone, E. (2011). *Recommendations for conducting differential item functioning (DIF) analyses for students with disabilities based on previous DIF studies* (ETS Research Report No. RR-11-34). ETS.
- Camilli, G. (1992). A conceptual analysis of differential item function in terms of a multidimensional item response model. *Applied Psychological Measurement*, 16(2), 129–147.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). American Council on Education/Praeger.
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–313.
- Campbell, D. T. (1979). *Assessing the impact of planned social change*. Public Affairs Center, Dartmouth College.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.
- Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple . . . *Language Testing*, 29(1), 19–27.

- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2007). *Building a validity argument for the test of English as a foreign language*. Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does and argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.
- Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multi-level construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Multi-level issues in organizational behavior and processes* (pp. 273–303). Elsevier.
- Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFLiBT scores to academic performance: Some evidence from American universities. *Language Testing*, 29(3), 421–442.
- Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods*, 17(1), 31–43.
- Cizek, G. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge: Taylor & Francis Group.
- Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4), 413–432.
- Clauser, B. E., Margolis, M. J., & Clauser, J. C. (2016). Issues in simulation-based assessment. In F. Drasgow (Ed.), *Technology and testing* (pp. 49–78). Routledge.
- Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives*, 9(4), 173–206.
- Cognia & Multi-State Alternate Assessment Consortium. (2019). *Multi-State Alternate Assessment 2019 technical report*. <https://www.azed.gov/assessment/technical-and-legal-resources>.
- Cole, M. (1996). *Cultural psychology: A once and future discipline*. Harvard University Press.
- Crisp, V. (2017). Exploring the relationship between validity and comparability in assessment. *London Review of Education*, 15(3), 523–535. <https://doi.org/10.18546/LRE.15.3.13>
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.
- Cronbach, L. L. (1976). Equity in selection—where psychometrics and political philosophy meet. *Journal of Educational Measurement*, 13(1), 31–41.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147–171). University of Illinois Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.
- Dadey, N., & Diggs, C. R. (2019). *A rapid review of interim assessment use*. The National Center for the Improvement of Educational Assessment.
- DePascale, C., & Gong, B. (2020). Comparability of individual students' scores on the "same test." In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability of large-scale educational assessments* (pp. 25–48). National Academy of Education.
- Deygers, B., van den Branden, K., & van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing*, 35(4), 449–476.
- Dolan, R. P., & Rose, D. (2000). Accurate assessment through universal design for learning. *Journal of Special Education Technology*, 15(4), 47–51.
- Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, 16(10), 640–647.
- Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Engelhard, G., Kobrin, J. L., & Wind, S. A. (2014). Exploring differential subgroup functioning on SAT writing items: What happens when English is not a test taker's best language? *International Journal of Testing*, 14(4), 339–359.
- Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Applications of think-aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice*, 29(2), 24–35.
- Ercikan, K., & Pellegrino, J. W. (2017). Validation of score meaning using examinee response processes for the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge.
- Ercikan, K., Roth, W., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, 27(4), 273–285.
- Ferrara, S. (2009). The Maryland School Performance Assessment Performance (MSPAP), 1991–2002: *Political Considerations*. Paper prepared for the Workshop of the Committee on Best Practices in State Assessment Systems: Improving Assessment While Revisiting Standards, December 10–11, National Research Council, Washington, DC. Available: http://www7.nationalacademies.org/BOTA/Steve_Ferrara_Paper.pdf
- Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research*, 103(2), 231–265.

- Gerasimova, D. (2024). Argument-based approach to validity: Developing a living document and incorporating preregistration. *Journal of Educational Measurement*, 61(2), 252–273. <https://doi.org/10.1111/jedm.12385>
- Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction*. CPRE Research Reports. https://repository.upenn.edu/cpre_researchreports/58
- Gordon, E. W. (1995). Toward an equitable system of educational assessment. *The Journal of Negro Education*, 64(3), 360–372.
- Gordon Commission. (2013). *To assess, to teach, to learn: A vision for the future of assessment* [Technical report]. ETS.
- Gould, S. J. (1981). *The mismeasure of man*. W. W. Norton & Company.
- Gregory, A., & Fergus, E. (2017). Social and emotional learning and equity in school discipline. *Future of Children*, 27(1), 117–136.
- Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6(4), 427–438.
- Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5(10), 511–517.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18, 4, 5–9.
- Haertel, E. (2004). Interpretive argument and validity argument for certification testing: Can we escape the need for psychological theory? *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 175–178.
- Haertel, E. H. (2018). Tests, test scores, and constructs. *Educational Psychologist*, 53(3), 203–216. <https://doi.org/10.1080/00461520.2018.1476868>
- Haertel, E., & Ho, A. (2016). Fairness using derived scores. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 217–236). Routledge.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Praeger.
- Harik, P., Clauser, B. E., Murray, C., Artman, C., Veneziano, A., & Margolis, M. (2013, April 26–30). *Comparison of automated scores derived from independent groups of content experts* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.
- Harwell, D. (2019, December). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *Washington Post*. <https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/>.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.
- Hood, S. (1998). Introduction and overview: Assessment in the context of culture and pedagogy: A collaborative effort, a meaningful goal. *The Journal of Negro Education*, 67(3), 184–186.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219–230.

- Hyman, J. (2017). ACT for all: The effect of mandatory college entrance exams on postsecondary attainment and choice. *Education Finance and Policy*, 12(3), 281–311.
- Hurwitz, M., Smith, J., Niu, S., & Howell, J. (2015). The Maine question: How is 4-year college enrollment affected by mandatory college entrance exams? *Educational Evaluation and Policy Analysis*, 37(1), 138–159.
- Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use this measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice*, 40(1), 92–100. <https://doi.org/10.1111/emip.12393>
- Jaeger, R. M. (1990). Establishing standards for teacher certification tests. *Educational Measurement, Issues and Practices*, 9(4), 15–20.
- Kachchaf, R., & Solano-Flores, G. (2012). Rater language background as a source of measurement error in testing of English language learners. *Applied Measurement in Education*, 25(2), 162–177.
- Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6, 125–160.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425–461.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135–170.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T. (2009). Certification testing as an illustration of argument-based validation. *Measurement*, 2(3), 135–170.
- Kane, M. T. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182.
- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3–17.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 198–211.
- Kane, M. (2017). Measurement error and bias in value-added models (ETS Research Report No. RR-17-25). ETS. <https://doi.org/10.1002/ets2.12153>
- Kane, M. T., Clauser, B. E., & Kane, J. (2017). A validation framework for credentialing tests. In S. Davis-Becker & C. W. Buckendahl (Eds), *Testing in the professions: Credentialing policies and practices* (pp. 21–40). Routledge.

- Kane, M. T., & Mislevy, R. (2017). Validating score interpretations based on responses processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). Routledge.
- Kelley, T. L. (1927). *Interpretation of educational measurement*. World Book Co.
- Keng, L., & Marion, S. (2020). Comparability of individual students' scores on the "same test." In A. Berman, E. Haertel, & J. Pellegrino (Eds.), *Comparability of large-scale educational assessments* (pp. 25–48). National Academy of Education.
- Kolen, M. J. (2011). *Comparability issues associated with assessment for the Common Core State Standards* [Paper presentation.] NCME Invited Symposium: Will Common Core State Standards and Assessment Consortia Ensure Comparability of Assessment Results Across Students, Schools, Districts, and States?, New Orleans, LA, United States.
- Konstantopoulos, S., Li, W., Miller, S. R., & van der Ploeg, A. (2016). Effects of interim assessments across the achievement distribution: Evidence from an experiment. *Educational and Psychological Measurement*, 76(4), 587–608. <https://doi.org/10.1177/0013164415606498n>
- Kopriva, R. J. & Wright, L. (2017). Score processes in assessing academic content of non-native speakers. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessment* (pp. 100–112). Routledge.
- Lane, S. (2014). Validity evidence based on testing consequences, *Psicothema*, 26(1), 127–135.
- Lane, S. (2017). Commentary II: The value of response process studies to evaluate score meaning. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessment* (pp. 138–144). Routledge.
- Lane, S. (2020). *Test-based accountability systems: The importance of paying attention to consequences* (ETS Research Report Series: 17th William H. Angoff Memorial Lecture). ETS.
- Lane, S., & DePascale, C. (2016), Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77–106). Routledge.
- Lane, S., & Iwatani, E. (2016). Design of performance assessments in education. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 274–293). Routledge.
- Lane, S., & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education*, 39, 165–215.
- Lane, S., & Niepokoj, D. (2018, April 12–17). Validity evidence based on consequences of assessment and accountability programs [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.
- Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 24–28.
- Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and

- student learning: Evidence from survey data and school performance. *Educational Assessment*, 8(4), 279–315.
- Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, 21(1), 23–30.
- Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement* (pp. 387–432). American Council on Education & Praeger.
- Leighton, J. P. (2017). Collecting and analyzing verbal response process data in the service of interpretive and validity arguments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 25–38). Routledge.
- Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A Festschrift for Frederick M. Lord* (pp. 27–40). Lawrence Erlbaum.
- Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15(1), 1–16.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 21(1), 14–16.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29(2), 4–16.
- Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195–212). Information Age Publishers.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of the requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 20(8), 15–21.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15–21.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635–694.
- Mao, L., Liu, O., Roohr, K., Belur, V., Mulholland, M., Lee, H., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, 23(2), 121–138. <https://doi.org/10.1080/10627197.2018.1427570>
- Marion, S. F. (2010). Constructing a validity argument for alternate assessments based on modified achievement standards. In M. Perie (Ed.), *Teaching and assessing low-achieving students with disabilities: A guide to alternate assessments based on modified achievement standards* (pp. 247–268). Brooks Publishing.
- Marion, S. F. (2018a). The opportunities and challenges of a system approach to assessment. *Educational Measurement: Issues and Practice*, 37(1), 45–46.
- Marion, S. F. (2018b). *What's wrong with high school testing and what can we do about it?* (Occasional Paper No. 29). National Center for the Improvement of Educational Assessment.

- Marion, S. F., & Domaleski, C. (2019). An argument in search of evidence: A response to Camara et al. (2019). *Educational Measurement: Issues and Practice*, 38(4), 27–28.
- Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, 25(4), 47–57.
- Marion, S. F., & Perie, M. (2009). Validity arguments for alternate assessments. In W. Schafer & R. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 115–127). Brooks Publishing.
- Marion, S., Thompson, J., Evans, C., Martineau, J. & Dadey, N. (2019, April 4–8) *A tricky balance: The challenges and opportunities of balanced systems of assessment* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, ON, Canada. <https://www.nciea.org/library/tricky-balance-challenges-and-opportunities-balanced-systems-assessment>
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 21(1), 16–18.
- Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist*, 20(2), 136–142.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35(11), 1012–1027.
- Messick, S. (1998, November). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment*. ETS. <https://doi.org/10.1002/j.2333-8504.1998.tb01797.x>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). American Council on Education and Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Michaels, H., & Ferrara, S. (1999). Evolution of educational reform in Maryland: Using data to drive state and policy local reform. In G. J. Cizek (Ed.), *Handbook of educational policy* (pp. 101–127). Academic Press.
- Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, 53(3), 265–292.
- Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4(1), 1–38.
- Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing* (pp. 23–48). Routledge.

- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice*, 23(2), 236–251. <https://doi.org/10.1080/0969594X.2015.1072085>.
- Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, 50(1), 91–98.
- National Research Council. (2001). *Knowing what students know*. National Academies Press.
- National Research Council. (2010). *Best practices for state assessment systems, Part I: Summary of a workshop*. The National Academies Press. <https://doi.org/10.17226/12906>
- National Research Council. (2017). *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress* (Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading, C. Edley & J. A. Koenig, Eds.). The National Academies Press. <https://doi.org/10.17226/23409>
- Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.
- Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10(1–2), 1–29.
- Newton, P. E. (2016). Macro- and micro-validation: Beyond the “five sources” framework for classifying validation evidence and analysis. *Practical Assessment, Research & Evaluation*, 21(12), 1–13.
- Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. SAGE.
- Newton, P. E., & Shaw, S. (2016a). Disagreement over the best way to use the word “validity” and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, 23(2), 178–197.
- Newton, P. E., & Shaw, S. D. (2016b). Agreements and disagreements over validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 316–318.
- Nisbet, I., & Shaw, S. D. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 26(5), 612–629.
- Nortvedt, G. A., Wiese, E., Brown, M., Burns, D., McNamara, G., O’Hara, J., Altrichter, H., Fellner, M., Herzog-Punzenberger, B., Nayir, F., Oya Taneri, P. O. (2020). Aiding culturally responsive assessment in schools in a globalising world. *Educational Assessment, Evaluation and Accountability*, 32(1), 5–27.
- Obama, B. (2014). *Remarks by the president at College Opportunity Summit December 4, 2014*. The White House, Office of the Press Secretary. <https://obamawhitehouse.archives.gov/the-press-office/2014/12/04/remarks-president-college-opportunity-summit>

- O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and uses of scores as aspects of validity. *Educational Measurement: Issues and Practice*, 36(2), 16–23. <https://doi.org/10.1111/emip.12141>
- Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, 27(4), 286–300.
- Oliveri, M. E., Lawless, R., & Mislevy, R. J. (2019). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. *International Journal of Testing*, 19(3), 270–300.
- Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analyzing, and interpreting response time, eye tracking and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). Routledge.
- Parke, C. S., & Lane, S. (2008). Examining alignment between state performance assessment and mathematics classroom activities. *Journal of Educational Research*, 101(3), 132–147.
- Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, 12(3), 239–269.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Chapter 9 : Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education*, 24(1), 307–353. <https://doi.org/10.3102/0091732X024001307>
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, 51(1), 59–81.
- Penuel, W. R., & Shepard, L. A. (2016a). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 787–850). AERA.
- Penuel, W. R., & Shepard, L. A. (2016b). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). John Wiley.
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13.
- Popham, W. J. (1997). Consequential validity: Right concerns-wrong concept. *Educational Measurement: Issues and Practice*, 21(1), 9–13.
- Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (1998). *Qualifying essay readers for an online scoring network (OSN)* (ETS Research Report No. RR 98-20). ETS.
- Prosser, R. & Solano-Flores, G. (2011, April 8–12). *Examining problem solving strategies on multiple-choice science items among English language learners through cognitive interviews* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States.
- Qualls, A. L. (1998). Culturally responsive assessment: Development strategies and validity. *The Journal of Negro Education*, 67(3), 296–301.

- Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C. W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, 105(4), 1100–1114.
- Quesen, S., & Lane, S. (2019). Differential item functioning for accommodated students with disabilities: Effect of differences in proficiency distributions. *Applied Measurement in Education*, 32(4), 337–349.
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(2), 82–90. <https://doi.org/10.1111/emip.12429>
- Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, 27(2), 170–178. <https://doi.org/10.1080/10627197.2022.2042682>
- Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.
- Rosen, Y. (2017). Assessing students in human-to-agent settings to inform collaborative problem-solving. *Journal of Educational Measurement*, 54(1), 36–53.
- Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.
- Russell, M. (2022). Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement: Issues and Practices*, 41(2), 25–35. <https://doi.org/10.1111/emip.12453>
- Ryan, K. (2002). Assessment validation in the context of high stakes assessments. *Educational Measurement: Issues and Practice*, 21(1), 7–15.
- Scoular, C., Care, E., & Hesse, F. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement*, 54(1), 12–35.
- Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, 19(2), 159–176.
- Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4(4), 447–467.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19(1), 405–450.
- Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 21(1), 5–8, 13.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14.
- Shepard, L. (2016). Evaluating test validity: Reprise and progress. *Assessment in Education: Principles, Policy, & Practice*, 23(2), 268–280.
- Shepard, L. A., Diaz-Bilello, E., Penuel, W. R., & Marion, S. F. (2020). *Classroom assessment principles to support teaching and learning*. Center for Assessment, Design, Research and Evaluation, University of Colorado Boulder.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34.

- Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45(1), 83–117.
- Sireci, S. G. (2009). No more excuses: New research on assessing students with disabilities. *Journal of Applied Testing Technology*, 10(2), 1–18. <http://www.testpublishers.org/assets/documents/Special%20Issue%20Article%201.%20pdf.pdf>
- Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50, 99–104.
- Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Bastera, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment: Addressing linguistic and cultural diversity* (pp. 3–21). Routledge.
- Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, 4, 43. <https://doi.org/10.3389/feduc.2019.00043>
- Solano-Flores, G., & Li, M. (2009). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
- Stone, C. A., & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education*, 16(1), 1–26.
- Stone, E., Laitusis, C. C., & Cook, L. L. (2016). Increasing the accessibility of assessments through technology. In F. Drasgow (Ed.), *Technology and testing* (pp. 217–234). Routledge.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). Universal design applied to large scale assessments (NCEO Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes.
- Thurstone, L. L. (1932). *The reliability and validity of tests*. Edwards Brothers.
- Thurstone, L. L. (1952). *The criterion problem in personality research* (Psychometric Lab Report, No. 78). University of Chicago.
- Trumbull E. & Nelson-Barber, S. (2019). The ongoing quest for culturally responsive assessment for Indigenous students in the U.S. *Front. Educ*, 4(40), 1–11.
- von Davier, A., Zhu, M., & Kyllonen, P. C. (2017). Introduction. In A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 1–18). Springer.
- Williamson, D. M., Bejar, I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, R. J. Mislevy, & I. Behar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 146–173). Routledge.
- Williamson, D. M., Xi, X., & Breyer, F. B. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.

- Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practices*, 37(1), 5–20.
- Winter, P., Kopriva, R. J., Chen, S., & Emick, J. (2006). Exploring individual and item factors that affect assessment validity for diverse learners: Results from a large-scale cognitive lab. *Learning and Individual Differences*, 16, 267–276.
- Wyse, A. E. (2011). The similarity of Bookmark cut scores with different response probability values. *Educational and Psychological Measurement*, 71, 963–985.
- Wyse, A. E. (2018). Regression effects in Angoff ratings: Examples from credentialing exams. *Applied Measurement in Education*, 31(1), 66–78.
- Zhai, X., Krajcik, J., & Pellegrino, J. (2021). On the validity of machine learning-based next generation science assessments: An inferential validity network. *Journal of Science Education and Technology*, 30, 298–321.
- Zwick, R. (2018). Fairness issues in the assessment of college and career readiness. In K. L. McClarty, K. D. Mattern, & M. N. Gaertner (Eds.), *Preparing students for college and careers* (pp. 94–103). Routledge.
- Zumbo, B. D., & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High stakes testing in education—science and practice in K–12 settings* (pp. 177–190). American Psychological Association Press.
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer.

NOTES

1. The *Standards* discussed fairness in terms of validity of interpretations for intended uses, whereas we argue that fairness pertains to validity of interpretations *and* uses *and* the resulting consequences.
2. Loevinger (1957) reported that a subcommittee (Meehl and Challman) of the APA committee drafted the first version of the *Technical Recommendations* and was the first to propose the term construct validity.
3. The 2014 *Standards* used the phrase *evidence for validity and consequences of testing*, as if to separate validity from consequential evidence.
4. Chapelle et al. (2010) used the term *utilization inference* instead of *decision/use inference*, which aligned with the terminology used by Bachman (2005).

Reliability in Educational Measurement

Won-Chan Lee

University of Iowa

Deborah J. Harris

University of Iowa (Retired)

This chapter is dedicated to exploring the concepts, quantification, and estimation of reliability and measurement errors, primarily within the framework of three prominent measurement theories: classical test theory (CTT), generalizability theory (GT), and item response theory (IRT). CTT, a well-established presence in the field of educational measurement, has consistently featured in previous editions of *Educational Measurement*. GT, having a shorter history than CTT, is often perceived as an extension or liberalization of CTT. The reliability chapters in prior two editions of *Educational Measurement* (Feldt & Brennan, 1989; Haertel, 2006) have provided relatively extensive treatments of GT. While the concept of reliability and measurement error can be discussed in the context of IRT, previous emphasis in IRT literature has typically centered around other aspects of the theory, such as model types, estimation of item and proficiency parameters, model fit, test construction, and scoring (see Hambleton, 1989; Yen & Fitzpatrick, 2006). With the growing applications of IRT in research and operational settings, it has become commonplace to present reliability information based on IRT models. Therefore, there is a pressing need to develop a cohesive framework for discussing reliability-related issues under the general assumptions of IRT. These three measurement theories offer distinct perspectives on various aspects of reliability, grounded fundamentally in their assumptions, which have significant consequences for the quantification and interpretation of reliability. This chapter primarily focuses on highlighting the similarities and differences among the theories in terms of conceptualizing and estimating various reliability statistics.

TERMS AND BACKGROUND

Measurement procedures are developed in accordance with the intended purposes of testing and the use of test scores for specific objects of measurement (e.g., test takers). This involves the identification of a set of tasks (e.g., multiple-choice items, essay prompts), administration modes (e.g., paper-pencil, computerized adaptive), scoring procedures (e.g., number correct, human rating), types of reported scores (e.g., summed raw scores, transformed scale scores), and score interpretations (e.g., norm-referenced, criterion-referenced). A specified set of measurement conditions constitutes a *test form*, with alternate forms encompassing different sets of similar (or parallel) measurement conditions. A sample of a test taker's responses to a test form is collected to generate one or more observed scores, which are then used to make general inferences about the test taker. The *observed score* for a test taker is considered a realization of a hypothetical distribution of all possible observed scores across repeated measurements using parallel forms of a measurement procedure. This observed score functions as an indicator or proxy for the underlying, unobservable *construct(s)* (i.e., domain of knowledge or skills) that a measurement procedure intends to measure. In IRT, proficiency estimates serve a similar role to observed scores.

Reliability, broadly conceived, is concerned with quantifying the extent to which observed scores are consistent over replications of a measurement procedure. However, the term *reliability* is often used to refer to *reliability coefficients*, which represent just one of many types of measures assessing the consistency of test scores. In this chapter, we make a clear distinction between these two terms. We use *reliability* to encompass a wide array of concepts and statistics used for quantifying the consistencies and/or inconsistencies in scores for individual test takers or groups of test takers. In a similar vein, the term *measurement error* is used broadly here to refer to both the sources of score inconsistencies and the summary statistics quantifying them for individual test takers or groups of test takers. It is evident that the characteristics of a measurement procedure play a crucial role in shaping the definition, interpretation, and estimation of statistics for reliability and measurement error. A particular focus will be placed on emphasizing the importance of reporting results for various statistics on the metric of *reported scores* used to make decisions about test takers. This is crucial because switching from one score scale to another, especially when the transformation is nonlinear, can lead to markedly different results and interpretations.

In his last paper, Cronbach (2004) stated,

In the history of psychometric theory, there was virtually no attention to this distinction [between sample and population] prior to 1951. . . . It was not until Lord's (1955) explicit formulation of the idea of random parallel tests that we began to write generally about the sampling, not only of persons, but of items. This two-way sampling had no counterpart in the usual thinking of psychologists. No change in procedures was required, but writing had to become more careful to recognize the sample–population distinction. (pp. 401–402)

This statement recognizes Lord's significant contribution in bringing attention to the concept of *replication*, particularly in terms of sampling items. The distinction highlighted here is between an estimate of reliability specific to a particular sample of items and the desired population value that would be obtained over many other random sets of such items. The concept of replications becomes a central element in modern theory of reliability (Brennan, 2001a; Feldt & Brennan, 1989; Haertel, 2006), and it is used as a unifying framework throughout this chapter for discussing various theories, models, and estimators of reliability. The sample–population distinction mentioned in the above quote naturally draws attention to two overarching notions or phases in reliability analysis: *conceptualization* and *estimation*. Brennan (2001a, 2001b, 2006) similarly emphasized this distinction, albeit in a slightly different manner.

The notion of *conceptualization* refers to an investigator's conception of the intended use and interpretation of reliability, which calls for defining a set of measurement conditions for replications of the measurement procedure. The set of conditions allowed to vary across replications are the sources of inconsistencies in scores (i.e., measurement error), which need to be considered when estimating reliability. It is important to note that there is no universally correct or optimal definition of replications.

The determination of which sources of measurement error to include is at the discretion of the investigator and depends primarily on the intended uses and interpretations of test scores.

The *estimation* phase involves quantifying reliability estimates (i.e., estimated reliability statistics) that best capture the characteristics defined in the conceptualization phase and requires decisions on the appropriate reliability statistics and data collection designs. The meaningfulness of interpretations from a reliability estimate hinges on how faithfully the specifications in the conceptualization phase are reflected in the estimation phase. Thus, it is essential to choose and compute a reliability estimator using an appropriate data collection design that effectively incorporates, within practical constraints, the influence of the specified sources of measurement error deemed important by the investigator. If the chosen reliability statistic involves sources of measurement error different from those conceptualized, an explanation is necessary. This is because the results could either overestimate or underestimate the parameter of interest, and understanding any divergence is crucial for accurate interpretation.

Consider, for example, an investigator wanting to estimate how consistent the rank orders of IRT proficiency estimates for candidates applying to a degree program remain when tested with a different set of multiple-choice items, possibly at different times. The measurement procedure involves the target population, exam characteristics, score metric, and score interpretation. The conceptualization envisions replication involving the use of different items and testing occasions.

For a direct estimation, data from two alternate test forms are collected from the same group at two separate occasions, applying the same IRT scoring method to both data sets. The estimated reliability coefficient is derived from the correlation between the proficiency estimates, aligning with the investigator's conceptualization of using different items and occasions.

Now, suppose the investigator chooses to report coefficient alpha based on a single administration with number-correct scores. This fails to address the original question, given that the metric of interest is IRT proficiency, not number-correct scores. Additionally, coefficient alpha overlooks testing occasions as a source of error, likely resulting in an underestimate of error variance and, consequently, an overestimate of the reliability coefficient. This example underscores the importance of selecting a reliability estimator that aligns with the conceptualized replications and is consistent with the intended use of the measurement procedure.

The preceding discussion on the conceptualization-estimation scheme is mostly aptly characterized within the framework of GT. However, the same principle is applicable to reliability analyses under CTT and IRT. The key distinction lies in the fact that, in CTT and IRT, investigators often select a reliability estimator and/or a data collection design without giving serious consideration to what constitutes replications. In effect, the investigator is making assumptions about replications, whether intentionally or unintentionally, through their choices. It should also be noted that

the three measurement models differ not only in their mathematical representations, but also in the framework for conceptualizing replications, true score, and measurement error, leading to potential ambiguities when interpreting results. The following section discusses some of the conceptual similarities and differences among the three measurement models regarding the definitions of true score and measurement error.

Throughout this chapter, the term *consistency* is primarily used to depict the variability of scores over replications. The term *precision* has also found use in the literature, including in the current *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA] et al., 2014), particularly in the context of reliability. Although the term precision literally denotes *exactness*, which diverges in meaning from the concept of consistency, it is occasionally used in this chapter when the context deems it more suitable. The term *accuracy* is specifically reserved for one of the classification indices, known as classification accuracy. This usage is retained to describe the degree of *correctness* of a measurement relative to the true value.

NOTIONS OF TRUE SCORE AND MEASUREMENT ERROR

Comprehending the concept of reliability requires grasping the notions of true score and measurement error. Almost all measurements in scientific disciplines involve observing the objects of measurement under certain conditions to produce *observed scores*. When observed under different conditions, the resulting observed scores are likely to differ, even if both measurements are intended to measure the same construct on the same objects of measurement. *True score*, by contrast, cannot be directly observed; rather, it is *defined*. True scores are pivotal in reliability, and how they are defined significantly affects the estimation and interpretation of reliability statistics. CTT and GT share similar perspectives on true score, while IRT employs a few distinct versions of true score.

In essence, true score in both CTT and GT is perceived as the *expected value* of observed scores over replications of a measurement procedure. This perspective precludes the platonic interpretation, which considers true score as a definitive indication of what the measurement intends to assess. The platonic view of true score has faced criticism because of its limited scientific utility, particularly in the context of psychological and educational measurement, where the constructs of interest are often intricate and challenging to explicitly define (see Lord & Novick, 1968, chap. 2).

The expected-value notion of true score requires specifying what constitutes a replication, implying that there is no singular definition of true score. The investigator has control over defining parallel forms for replications, such as different sets of items, raters, and/or occasions. Many well-known results in CTT rely on the assumption that replications are performed over forms that are *classically parallel*.

In contrast, later developments of single-administration reliability coefficients use less stringent assumptions, such as essentially tau-equivalent or congeneric forms. Meanwhile, true score in GT (also referred to as universe score) is defined as the expected value of observed scores over measurement conditions or forms that are *randomly parallel*. This distinct conception of parallelism for replications stands out as a key difference between the two theories, carrying considerable theoretical and practical implications.

The term *measurement error* refers to the discrepancy between a test taker's true score (i.e., the average of observed scores over replications) and the actual observed score. This implies that the definition of measurement error depends on how true score is conceptualized. Measurement error arises due to the random variation of observed scores over repeated testing using specified conditions of a measurement, which are used to define true score. Other error sources, deemed less important and thus not explicitly modeled by the investigator, are assumed to operate randomly. In CTT and GT, it is assumed that specified and unspecified (i.e., residual) sources of measurement error behave randomly such that their expected value over replications is always zero. If there are factors unrelated to the construct the test aims to measure that influence test-taker scores in some *systematic* way (e.g., verbal skills in math computation, bias in raters, model misfit), they are considered sources of construct-irrelevant variance and are deemed threats to *validity* (Kane, 2006). Rarely does construct-irrelevant variance play any role in estimating reliability under CTT or GT. However, it does hold implications in the context of IRT, as discussed next.

In unidimensional IRT, the notion of true score is conceptually linked to the person proficiency parameter, denoted θ . Although the proficiency parameter is explicitly defined and specific to a particular IRT model, subsequent discussions are intended to be general, applicable to any model. The proficiency parameter θ represents a person's location in a latent trait space and plays a role similar to that of true score. Consider a biased estimator of θ , such as the maximum likelihood estimator. Due to this bias, the expected value of the estimates across replications, denoted $\mathcal{E}\hat{\theta}$, does not equal the parameter θ —that is, $\mathcal{E}\hat{\theta} \neq \theta$. Unlike CTT, which adopts the expected-value notion of true score, the two possible definitions of a person's proficiency in IRT (i.e., θ and $\mathcal{E}\hat{\theta}$) contribute to some inconsistencies and ambiguities in estimating and interpreting reliability coefficients for IRT proficiency estimates. In this chapter, θ is referred to as *latent proficiency*, while $\mathcal{E}\hat{\theta}$ denotes *expected proficiency*, the latter being considered a more suitable definition in the context of reliability. Clear distinctions between these two definitions are not always made, and both have been used in various formulas for estimating reliability in IRT.

Different definitions of true proficiency inevitably result in distinct definitions of measurement error. Error of measurement, defined by the discrepancy between $\hat{\theta}$ and $\mathcal{E}\hat{\theta}$, constitutes random error, and its variance is referred to here as the *expected error variance*. By contrast, the variance of discrepancies between $\hat{\theta}$ and θ —referred to

simply as *error variance* here—involves both systematic error (i.e., bias) and random error. While bias can diminish validity, it does not impact reliability. If the variance (and covariance) of bias is incorporated into a reliability coefficient, it should contribute to true score variance rather than error variance, thereby increasing reliability. In practice, there is often an argument that the bias in proficiency estimates, particularly its variance, is negligible, making the distinction unnecessary. However, empirical examples presented by W. Lee et al. (2025) demonstrate that differences can sometimes be substantial.

In IRT, the definition of true score expressed in the number-correct (or summed raw) score metric is represented by a test characteristic curve (TCC). The TCC is a nonlinear transformation of θ , where values of θ with an unbounded range from $-\infty$ to ∞ are transformed into a score range of 0 to the maximum possible summed raw score. A TCC value represents the expected value of the model-based distribution of observed scores for a given θ , assuming known item parameters. Measurement error is characterized by the variability of model-based observed scores for a given θ . The TCC definition of true score in IRT is akin to true score in CTT in two key aspects: (a) It is expressed on the summed raw score metric, and (b) it employs the expected-value notion of true score as opposed to the latent-trait definition. Consequently, most results derived in CTT remain applicable for estimating reliability on the metric of summed raw scores in IRT. It is crucial to note that, similar to $\hat{\theta}$ (and θ), a TCC is model-specific, signifying that it is meaningful and can be interpreted properly only under the chosen IRT model.

The consideration that a TCC for a given model is often deemed fixed implies that the expectation is taken over forms containing a set of items with identical item parameters. This argument also extends to the expected proficiency, $\mathcal{E}\hat{\theta}$. Forms with identical item parameters are termed *strictly parallel* forms, effectively implying a fixed form. As a result, reliability coefficients under IRT should be larger than those based on CTT or GT, all other factors being equal.

TYPES OF SCORES CONSIDERED

Scores derived from test-taker responses can manifest in various forms, and the choice of a reporting metric introduces distinct considerations for reliability analyses. In this chapter, our focus centers on specific types of scores:

- *Summed raw scores*: These scores result from the summation of points earned on each individual test item. Variants include percent correct scores and weighted sum scores, where items may be assigned different weights based on content importance or other considerations. Summed raw scores may be referred to simply as *raw scores* in this chapter.
- *IRT proficiency estimates*: Derived from IRT models, proficiency estimates are based on test takers' responses and item parameters. While profi-

ciency estimates are not commonly used for reporting because of their interpretational complexity, they find frequent application in adaptive testing to guide item selection. Two different IRT proficiency estimators are considered in this chapter based on the maximum likelihood and Bayesian estimation methods.

- *Composite scores:* Composite scores are linearly weighted sums of component scores, either subscores from one test or scores from different tests. Multiple components can be characterized by the differences in content areas, constructs measured, item formats, and other aspects. Unlike simple summed raw scores, composite scores treat items across components distinctly.
- *Scale scores:* Widely employed for reporting, scale scores can be crafted in various ways, such as setting target means and standard deviations, incorporating precision information for equal conditional standard errors of measurement, or setting a cut score with a particular scale score value. Scale scores prove particularly useful in reporting when test takers take different sets of items. A conversion table is typically created to convert summed raw scores, IRT proficiency estimates, or composite scores to scale scores.
- *Classification category scores:* Instead of providing a specific numerical score, classification category scores convey a test taker's performance category, such as *Exceeds Expectations*, *Meets Expectations*, or *Below Expectations*. *Pass/Fail* or *Master/Nonmaster* are also common in various assessments.

STATISTICS AND INDICES FOR RELIABILITY

Multiple methods exist for quantifying reliability for individual scores or scores for a group of test takers, and some of these are defined next.

Overall Standard Error of Measurement

The overall standard error of measurement (SEM) represents the standard deviation, over individuals, of observed scores minus true scores for an assessment. Traditionally, it is estimated as a function of a reliability coefficient and observed score variance. This estimation is referred to as the overall SEM, distinguishing it from the conditional SEM. SEMs, both overall and conditional, are expressed in the same units as the reported score, making them specific to the scoring metric used. Consequently, SEMs cannot be directly compared across different scoring procedures. Another notable characteristic of SEM is its relative insensitivity to the characteristics of a specific group of test takers.

Conditional Standard Errors of Measurement

The SEM typically varies across score levels or test takers. The conditional SEM (CSEM), in theory, represents the standard deviation of observed scores over repeated

measurements conditional on a test taker's true score. In practice, estimated CSEMs can be reported either at each true score level based on a psychometric model or for each individual test taker using an observed score as an estimate of the true score. CSEMs provide valuable information about the amount of measurement error for each test taker (or score level) and can also be aggregated over all score levels to compute the overall SEM and reliability coefficients. The section "Estimators of CSEMs" of this chapter is dedicated to discussions on estimators of CSEMs for various types of scores.

Reliability Coefficients

The extent to which test takers' scores are consistent over replications can be quantified using reliability coefficients. These coefficients, originally developed under the traditional assumptions of CTT, take various forms. One such coefficient, for example, involves the correlation between observed scores on the same assessment or parallel forms of an assessment. Reliability coefficients are not reported in score units, making them challenging for users to interpret directly. Reliability coefficients are sensitive to the characteristics of the test-taker group. Some estimators require at least two scores per test taker, which is a challenging requirement to meet in practice, while others can be computed based on a single administration of an assessment. Various approaches to estimating reliability coefficients under each of the three measurement model frameworks are discussed in this chapter.

Classification Consistency and Accuracy Indices

When test scores are used to categorize test takers based on one or more cut scores, a crucial consideration is the likelihood of consistent classification if the test is administered again. Classification consistency serves as a criterion-referenced measure of reliability, assessing whether test-taker performance aligns with established standards or cut scores across replications. In contrast, classification accuracy is concerned with whether a test taker is "accurately" categorized into the performance category corresponding to their true score. This definition implies a closer connection to validity. Because consistency and accuracy are often considered together, both aspects are addressed in this chapter. The section "Reliability of Classification Category Scores" of this chapter is dedicated to presenting methods for estimating various classification consistency and accuracy indices.

ORGANIZATION OF THIS CHAPTER

This chapter builds on concepts discussed in earlier editions of *Educational Measurement*, particularly in Haertel (2006) and Feldt and Brennan (1989). The sections "Reliability in CTT" and "Reliability in GT" delve into classical and generalizability theory approaches. The section "Reliability in IRT" treats IRT approaches to reliability issues, offering a more extensive discussion because of the relatively less explored nature of this area. Some content discussed in the IRT section is novel and has not been previously published. Emphasizing the utility of CSEMs over reliability coefficients,

the section “Estimators of CSEMs” provides various estimation procedures for CSEMs. The section “Reliability of Classification Category Scores” focuses on reliability of classifications, while the section “Other Models, Aggregation, and Precision Issues” tackles various additional issues associated with reliability. The final section offers a concise summary of the chapter and outlines potential areas for future research on reliability. For practical applications and comparisons of different methodologies, readers can refer to two real data examples presented by W. Lee et al. (2025), which demonstrate the computation of various reliability and error statistics across different models and metrics.

RELIABILITY IN CTT

CTT emerged in the early 20th century, initially focusing on studies of individual differences. Its origins are often traced back to Spearman’s (1904) work on methods for correcting correlation coefficients for attenuation due to measurement error (Traub, 1997). Since then, numerous scholars have contributed to further developing and refining the theory. Key texts in the field, such as Gulliksen (1950) and Lord and Novick (1968), are considered highly influential, offering comprehensive treatments of CTT.

Despite its relatively simple set of assumptions and definitions compared to other contemporary theories, CTT yields results that are relevant to many aspects of modern psychometric applications and remains widely used. Not surprisingly, CTT continues to be an important area of study. In recent decades, certain segments of the measurement literature have extended CTT substantially, particularly in areas such as estimating CSEMs for both raw and scale scores, classification consistency and accuracy, and interval estimation (although the latter is not extensively covered in this chapter).

Assumptions, Definitions, and Basic Results

CTT asserts that an observed score for person p on form f , denoted X_{pf} , can be decomposed into two components: a true-score component, T_p , and an error-score component, E_{pf} :

$$X_{pf} = T_p + E_{pf}, \quad (1)$$

where the true score T_p is a constant specific to the person and does not depend on forms. For notational convenience, T_p and τ_p are used interchangeably in this chapter to represent the true score for person p in the total raw-score metric. The intrinsic difficulty of using the model in Equation 1 arises from the fact that only one variable can be observed, while the other two are unobservable. This challenge is circumvented by making certain assumptions about the unobservable true or error scores—assuming either one of them to be known makes the other apparent. One such assumption is that,

for a specific person p , the expected value ($\mathcal{E}$) of error scores over replications is zero: $\mathcal{E}_f(E_{pf}) = 0$, where the expected value is taken over a hypothetically infinite number of parallel forms, as indicated by the subscript f below the expectation operator. Under this assumption, it follows that

$$\mathcal{E}_f(X_{pf}) = \mathcal{E}_f(T_p + E_{pf}) = T_p. \quad (2)$$

Conversely, if the true score is defined as the expected value of observed scores, the expected value of errors is necessarily zero. Therefore, the true score in CTT, which is equal to the expected value of observed scores, can be viewed either as a derived result or as a definition.

Assuming the measurement procedure does not alter a person's true score, any variation in the observed scores over replications is attributed solely to the use of parallel, yet different forms of a test. The question of what constitutes the replications over which the expectations in Equation 2 are taken lies at the core of CTT because the nature of true score and measurement error depends on the answer to that question. One traditional answer, among many possibilities, is that expectations are taken over forms of a test that are *classically parallel*. However, as discussed later, different definitions of replications can lead to different results. It should be noted that there is no universally right or best definition of replications, and the choice depends entirely on the investigator's decision, guided by the intended interpretations of test scores.

Another assumption about error scores is $\mathcal{E}_p(E_{pf}) = 0$, meaning that, for a given test form, the expected value of the errors taken over a population of persons equals zero. This assumption holds for any subpopulation of persons, unless persons are selected based on the magnitude of X_{pf} . Under this and other assumptions discussed previously, the following central results can be derived: (a) true scores and errors for any test form are uncorrelated; and (b) error scores on any pair of parallel forms, f and g , are uncorrelated. It is also true that their covariances are zero, notationally, $\sigma(T, E_f) = 0$ and $\sigma(E_f, E_g) = 0$. It follows that the variance (over persons) of observed scores on a form is simply the sum of true score variance and error score variance:

$$\sigma^2(X_f) = \sigma^2(T) + \sigma^2(E_f). \quad (3)$$

It can be further shown that, under CTT assumptions, true score variance is equal to the covariance between observed scores on a pair of forms and also to the covariance between observed scores and true scores: $\sigma(X_f, X_g) = \sigma(X_f, T) = \sigma^2(T)$.

Reliability Coefficients and SEM

The historical definition of the reliability coefficient is rooted in the conceptually straightforward notion of replication, specifically the correlation between two parallel measurements. The notation $\rho(X_f, X_g) \equiv \rho(X, X')$ denotes the reliability coefficient,

defined as the correlation between observed scores arising from the same form or two parallel forms. Under the CTT assumptions and results set forth previously, other expressions of the reliability coefficient can be derived (note that, without loss of generality, the subscript f for test form will be dropped hereafter):

$$\rho(X, X') = \rho^2(X, T) = \frac{\sigma^2(T)}{\sigma^2(X)} = 1 - \frac{\sigma^2(E)}{\sigma^2(X)}. \quad (4)$$

It is important to note that these equalities are derived under specific assumptions in CTT, particularly those pertaining to classically parallel forms, and should not be universally applied without consideration. For instance, as detailed later, alternative expressions of the reliability coefficient may produce different results in certain IRT contexts. Additionally, in this chapter, the terms *definitions* and *expressions* are used interchangeably to offer clearer distinction in various situations.

Using $R(X)$ as a generic notation for the reliability coefficient in the raw-score metric X , the following symbols distinguish various expressions presented in Equation (4):

$$R(X)_{PF} \equiv \rho(X, X')$$

$$R(X)_{SC} \equiv \rho^2(X, T)$$

$$R(X)_{VR} \equiv \sigma^2(T)/\sigma^2(X)$$

$$R(X)_{EV} \equiv 1 - [\sigma^2(E)/\sigma^2(X)]$$

This distinction gains increased importance and utility in IRT contexts. The parallel-forms definition, $R(X)_{PF} \equiv \rho(X, X')$, stands apart from the others by being expressed solely in terms of observable quantities without referencing the true or error component. One limitation of this definition is the absence of an explicit parameter for the reliability coefficient because the correlation may vary across different pairs of forms unless forms are definitely classically parallel.

In the extant literature, the squared-correlation definition, $R(X)_{SC} \equiv \rho^2(X, T)$, is regarded as the canonical definition of the reliability coefficient (Brennan, 2010). The generalizability coefficient in GT, which is analogous to the reliability coefficient in CTT, takes the form of the squared correlation. The variance-ratio definition, $R(X)_{VR} \equiv \sigma^2(T)/\sigma^2(X)$, employs the ratio of true score variance to observed score variance, indicating that if the quantity is close to 1.0, variability in test takers' observed scores is largely attributable to variability in their true scores rather than errors. This definition forms the basis for the development of many internal consistency reliability coefficients (e.g., coefficient alpha), where a key objective is to express unobservable true score variance in terms of some observable quantities. The last definition involving error variance, $R(X)_{EV} \equiv 1 - [\sigma^2(E)/\sigma^2(X)]$, always yields the same result as the variance-ratio

definition, as long as Equation 3 holds. This definition may be considered the most general form of the reliability coefficient because of its broad applicability to various score types.

The reliability coefficient essentially represents a correlation ranging from zero to 1, which is a convenient unitless scale for comparing different instruments or measurement procedures. By contrast, SEM is expressed in units of a reported score scale and describes the extent to which test takers' scores vary from one testing to another. Traditionally, SEM, in the raw-score metric, is derived through the reliability coefficient as

$$\sigma(E) = \sigma(X)\sqrt{1 - R(X)}. \quad (5)$$

Strictly speaking, this derivation holds only if the assumptions of classically parallel forms are met. However, in practice, these assumptions are often disregarded and loosely applied in conjunction with various reliability coefficient estimates. The consequence of violating these assumptions is relatively unknown.

From the perspective of CTT, the SEM derived in Equation 5 is termed the *overall* SEM for a population of persons. Alternatively, the overall SEM can be defined as the square root of the expected value of individual-level error variances, which typically differ across persons with different true score levels. Symbolically,

$$\sigma(E) = \sqrt{\mathcal{E}_p \sigma^2(E | \tau_p)} = \sqrt{\mathcal{E}_p \sigma^2(X | \tau_p)}, \quad (6)$$

where $\sigma^2(E | \tau_p) = \sigma^2(X | \tau_p)$ is called the *conditional* error variance for persons with fixed true score τ_p , and the square root of it is the CSEM. Numerous theoretical and empirical studies have highlighted that CSEMs exhibit variation along the score scale, raising concerns about the applicability of the overall SEM to all individuals. (Refer to the section "Estimators of CSEMs" for estimators of CSEMs for various score types under all three measurement models.) It is important to note that the exact mathematical relationship between the overall and conditional SEMs in Equation 6 holds only for certain models and estimators.

In the measurement literature, reliability coefficients have been the central focus, with insufficient attention given to the role of CSEMs. CSEMs not only provide information about the amount of measurement error specific to each individual or score point but also can be used to estimate overall statistics such as the overall SEM and reliability coefficients. They can further be integrated into interval estimation procedures (K. Y. Kim & Lee, 2018; W. Lee et al., 2006). In addition, obtaining CSEMs for scale scores is relatively straightforward.

Approaches to Estimating Reliability

Approaches to estimating the reliability coefficient and SEM in CTT can be broadly categorized into two types. In the first type, reliability is directly estimated using data from two independent replications of the "full-length" measurement procedure

administered to a group of test takers. The resulting two sets of observed scores at the group or individual levels are used for computing reliability coefficients, SEM, CSEMs, and other relevant statistics. In contrast, the second type estimates reliability using data collected from a single test administration to avoid the challenges of conducting two actual replications. Most estimators of the second type involve splitting a full-length test form into two or more constituent part-tests. The data from these part-tests from a single group of test takers are treated as replications—a somewhat contrived notion of replications. Statistical quantities derived from part-tests are then used to obtain results for the full-length test form. However, this approach relies on a strong set of assumptions regarding the parallelism of the part-tests.

In principle, data for a reliability analysis are collected in a way that properly reflects the influences of all important sources of error in the reliability statistics of interest. To generalize interpretations of test results beyond a particular incident of a measurement procedure, investigators would need to conceptualize which measurement conditions shall vary over replications and design data collection accordingly. Ideally, in the data collection process, only the measurement conditions contributing to the conceptualized measurement error should be allowed to vary over replications, while all other potential sources of random error remain unchanged or are treated as negligible.

Direct Estimation With Full-Length Replications

Suppose an investigator is interested in generalizing test takers' test scores beyond a particular collection of items on a test form. A straightforward approach is to construct two full-length parallel forms of the test according to the same test specifications. These two forms are then administered (possibly in a counterbalanced manner) to the same group of test takers within a very short time interval, such as on the same day. The correlation between two sets of scores is subsequently calculated. This correlation value serves as a direct estimate of the reliability coefficient for either test form (not the sum or average of the two). Known as a *coefficient of equivalence*, this correlation indicates the extent to which test takers perform similarly on different forms of the test.

When two test forms are administered at different times (e.g., a few days apart), the resulting correlation is affected downward due to the measurement error arising from differences in both testing occasions and test forms. This correlation is referred to as a *coefficient of stability and equivalence*, which is often preferred over other estimates because investigators typically seek to generalize results over different testing occasions as well as different test forms. In this case, the coefficient of equivalence might be an overestimate of the idealized (or conceptualized) reliability. Alternatively, in situations where only one test form is available and administered twice, typically at different times, the estimate is called a *coefficient of stability* or *test-retest reliability coefficient*. This coefficient reflects variability in observed scores over testing occasions,

which may be biased by memory, but it does not account for variability due to different test forms. It is important to emphasize that these coefficients are not interchangeable principally because they involve the impact of different sources of measurement error. The choice of which coefficient to use depends on its consistency with the intended generalization.

The direct estimation approaches using actual full-length replications offer several advantages. First, in theory, they can be applied to any measurement procedure, including performance assessments involving raters, as long as resources permit. Second, these approaches require minimal statistical and psychometric assumptions, although they do rely on other assumptions such as no change in true score or latent traits between administrations, no memory effects on the second responses, and no fatigue or practice effects. Third, the computation is easy and straightforward. Fourth, other reliability statistics, including CSEMs, can be readily computed from the replicated data. Fifth, it provides the investigator an opportunity to carefully consider what constitutes replications of the measurement procedure, with a central focus on the sources of measurement error (e.g., items, occasions, raters). Despite all these benefits, however, direct estimates are less frequently used in practice because of their requirement for double testing time and, for some coefficients, the construction of an additional parallel form.

Estimation Based on a Single Test Administration

Spearman (1910) and Brown (1910) were the first to recognize the necessity of estimating a reliability coefficient using data from a single test administration. Since then, numerous estimators have been developed, collectively known as *internal consistency coefficients*. The well-known Spearman–Brown formula was developed for the purpose of estimating a reliability coefficient based on two split-halves with equal observed score means and variances (i.e., classically parallel). Recognizing the challenge of constructing part-tests with precisely identical means and variances, subsequent developments attempted to relax the strict assumptions of classically parallel forms. In particular, Feldt and Brennan (1989) integrated various internal consistency coefficients and categorized them based on varying degrees of part-test parallelism.

The notions of tau-equivalent and essentially tau-equivalent forms were initially introduced by Lord and Novick (1968). For two forms (or part-tests) that are *tau equivalent*, it is assumed that true scores are the same, thereby necessitating equal observed score means. However, these two forms are allowed to have different error variances, leading to the possibility of differing observed score variances. Under *essentially tau-equivalent* forms assumptions, the requirement of equal observed score means can be relaxed further. Here, true scores may differ by a constant, allowing for variations in observed score means. Despite these differences, true score variances are equal for essentially tau-equivalent forms due to the constant difference in true scores. A less stringent concept of parallelism is *congeneric* forms, where true

scores are assumed to have a linear relationship. Consequently, true score variances may vary, but they are perfectly correlated. Under the congeneric assumptions, observed score variances may also differ due to variations in both error and true score variances.

Extensive discussions and derivations of internal consistency coefficients, considering various definitions of form parallelism, can be found in Feldt and Brennan (1989) and Haertel (2006). In this section, a subset of these coefficients will be briefly introduced, starting with those based on two parts and then extending to coefficients involving more than two parts.

For two-part coefficients, the full-length test X is divided into two scorable part-tests, X_1 and X_2 , such that $X = X_1 + X_2$. The Spearman–Brown formula, the first internal consistency coefficient in the history of measurement, was derived under the assumption that X_1 and X_2 are classically parallel. The correlation between the two part-tests, denoted by $\rho(X_1, X_2)$, serves as a reliability coefficient for a half-test. This coefficient is then stepped up to obtain a reliability coefficient for the full-length test X using the simple formula

$$_{SB}R(X) = \frac{2\rho(X_1, X_2)}{1 + \rho(X_1, X_2)}. \quad (7)$$

Other two-part coefficients developed subsequently relax the classically parallel form assumptions to express the unobservable true score variance in terms of observable quantities based on two parts. Rulon (1939) provided a formula (attributed to Flanagan) under the assumption of essential tau equivalence, where the true score variance for the total test is expressed to be equal to four times the covariance between X_1 and X_2 : $\sigma^2(T_X) = 4\sigma(X_1, X_2)$. Algebraically identical versions were proposed by Rulon (1939) and Guttman (1945). Putting all three versions together, the Flanagan–Guttman–Rulon coefficient is

$$_{FGR}R(X) = \frac{4\sigma(X_1, X_2)}{\sigma^2(X)} = 2 \left(1 - \frac{\sigma^2(X_1) + \sigma^2(X_2)}{\sigma^2(X)} \right) = 1 - \frac{\sigma^2(X_1 - X_2)}{\sigma^2(X)}. \quad (8)$$

The assumption of equal true score variance in essential tau equivalence is rarely satisfied when, for example, two parts have substantially different test lengths or are associated with different content areas. Allowing for different part-test true score variances increases the number of unknown quantities and calls for additional constraints to find a solution. Raju (1970) introduced a constraint that the relative lengths of the two parts, l_1 and $l_2 = 1 - l_1$, are known (e.g., the actual number of items in each part-test). In contrast, Angoff (1953) and Feldt (1975) independently proposed the same solution for situations in which the lengths of part-tests are treated as unknown parameters to be estimated. They used a constraint that the part-test error variances be linearly related to each other as a function of the l terms. It follows that the unknown part-test lengths are equal to *effective test lengths*: $l_1 = [\sigma^2(X_1) + \sigma(X_1, X_2)] / \sigma^2(X)$ and $l_2 = [\sigma^2(X_2) + \sigma(X_1, X_2)] / \sigma^2(X)$. The Angoff–Feldt coefficient is given by

$${}_{AF}R(X) = \frac{\sigma(X_1, X_2)}{l_1 l_2 \sigma^2(X)} = \frac{4\sigma(X_1, X_2)}{\sigma^2(X) - \left(\frac{\sigma^2(X_1) - \sigma^2(X_2)}{\sigma(X)} \right)^2}. \quad (9)$$

Note that Raju's coefficient is a special case of Equation 9 where l_1 and l_2 are predefined constants.

Among many possible methods for splitting a test into two scorable parts, the odd-even split is most frequently employed in practice. The goal in dividing a test is to create two half-tests, each of which reflects the specifications of the full-length test and is as similar as possible to each other in content and item characteristics (e.g., difficulty, format) to closely adhere to parallelism assumptions. Failure to achieve this similarity may introduce bias into estimates of reliability. An inherent limitation of the two-part approaches is that, regardless of the split-half strategy used, there are many possible split halves that are equally acceptable. However, not all such splits will produce the same estimate.

Subsequently, the two-part approaches were extended to multipart. Multipart estimates, often derived from individual items treated as parts, are generally preferred over two-part estimates because they avoid arbitrary divisions of a test, and the sampling errors in reliability coefficients tend to be smaller (Kristof, 1963). However, two-part approaches may be more defensible than multipart approaches concerning parallel-part assumptions. For example, the claim that all items are essentially tau equivalent is rarely justifiable because it implies that any differences in the item-level observed score variances are solely due to measurement error. Constructing two parts that largely satisfy essential tau equivalence might be much easier than developing all individual items with such a high degree of parallelism. Therefore, multipart procedures should not necessarily be perceived as "better" than two-part procedures.

Suppose a full-length test X can be divided into n parallel part-tests: $X = X_1 + X_2 + \cdots + X_n$, where each part-test can be an individual item or a cluster of items. Under the assumption of classically parallel part-tests, the generalized Spearman-Brown (GSB) formula can be derived as

$${}_{GSB}R(X) = \frac{nR(X_1)}{1 + (n - 1)R(X_1)}, \quad (10)$$

where $R(X_1)$ is a reliability coefficient for a unit-length test (i.e., a part-test). If the above equation is solved for $R(X_1)$, one can obtain a prophecy formula for predicting the reliability of a shortened test. Therefore, a more general formula can be written as

$${}_{pred}R(X) = \frac{v \times_{orig} R(X)}{1 + (v - 1) \times_{orig} R(X)}, \quad (11)$$

where ${}_{orig}R(X)$ is the reliability coefficient of the original test, and v is the ratio of the length of the predicted test to the length of the original test. Obviously, v can be any real number greater than zero.

Cronbach's (1951) alpha is undeniably one of the most widely recognized multipart reliability coefficients, assuming essentially tau-equivalent part-tests.¹ The formula is given by

$${}_aR(X) = \left(\frac{n}{n-1} \right) \frac{\sigma^2(X) - \sum \sigma^2(X_i)}{\sigma^2(X)}, \quad (12)$$

where $\sigma^2(X_i)$ is the variance of each part-test, $X_i (i = 1, \dots, n)$. Note that the multiplier $n/(n-1)$ has nothing to do with correcting for bias. Notably, coefficient alpha is algebraically equal to Kuder and Richardson's (1937) KR20 coefficient if each item is treated as a part-test, and all items are scored dichotomously (i.e., 0 or 1). Kuder and Richardson's KR21 is a special case of KR20 when items are all equally difficult. However, because of the unrealistic assumption of equal item difficulty, the practical utility of KR21 is limited.² Another well-known property of ${}_aR(X)$ is that it equals the average of all possible split-half reliability coefficients computed using ${}_{FGR}R(X)$ in Equation 8. As a special case, when $n = 2$ in Equation 12, ${}_aR(X) = {}_{FGR}R(X)$.

Reporting coefficient alpha is prevalent and nearly a standard practice in the measurement and assessment literature. However, it is also true that alpha is frequently misused and/or misunderstood. One example of improper use of coefficient alpha is to assess the dimensionality of items in a test. Nothing in the derivation of alpha requires unidimensionality—alpha is not based on a latent trait model. In his reflection on coefficient alpha 50 years later, Cronbach (2004) stated, "I particularly cleared the air by getting rid of the assumption that the items of a test were unidimensional" (p. 397).

Another widely cited but potentially misleading characteristic of alpha is the phrase that alpha is a lower limit to reliability (Lord & Novick, 1968). While the mathematical proof is valid under a particular set of assumptions, it is hardly generalizable to other measurement circumstances for which the sources of error are different from those involved in alpha. For example, if the investigator intends to generalize over different testing occasions in addition to different forms, coefficient alpha computed based on data from a single occasion will likely be an overestimate (Brennan, 2001b, 2010). As mentioned earlier, the crux of the matter is whether the data used for estimation reflect the effects of error sources to be consistent with the investigator's conceptualization of replications. Cronbach (2004) also stated, "I no longer regard the alpha formula as the most appropriate way to examine most data" (p. 403). In theory, depending on the investigator's conceptualization, there can be many reliability coefficients for any set of test scores, and Cronbach incorporated this notion into the development of GT.

Another coefficient for essentially tau-equivalent parts, developed by Guttman (1945), is coefficient λ_2 , which is one of Guttman's series of lower bounds for reliability referred to as λ_1 through λ_6 . Guttman's λ_2 is always greater than or equal to λ_3 , which is the same quantity as alpha, but is lower than "the" reliability. In this sense, λ_2 is often considered a "better" lower bound than alpha. However, again, the conceptual

definition of reliability with respect to sources of error for replications should take precedence over any statistical justification for lower bounds.

If the stringent assumptions of essentially tau-equivalent parts are relaxed to be congeneric, part-tests are allowed to have heterogeneous error variances and true score variances, while maintaining a perfect correlation between true scores. The congeneric model is deemed suitable for cases in which observed part-test variances differ considerably, and one cannot solely attribute such variations to measurement error. For example, consider a test comprising several units (e.g., reading passages), each associated with a different number of items. If these units serve as part-tests in a reliability analysis, the observed part-test variances may well differ, and the differences are likely attributable not only to measurement error but also to variations in true score variances.

Under the congeneric assumptions, Raju (1977) extended his two-part coefficient (Raju, 1970) to multipart contexts, assuming the proportions of total test length for part-tests, l_i , are known:

$${}_R R(X) = \left[\frac{1}{1 - \sum l_i^2} \right] \left[\frac{\sigma^2(X) - \sum \sigma^2(X_i)}{\sigma^2(X)} \right] \quad (13)$$

which becomes identical to coefficient alpha if $l_i = 1/n$. For unknown part-test lengths, Feldt and Brennan (1989) provided a solution called Feldt's coefficient:

$${}_F R(X) = \frac{\sigma^2(X) [\sigma^2(X) - \sum \sigma^2(X_i)]}{[\sigma^2(X)]^2 - \sum [\sigma(X_i, X)]^2} \quad (14)$$

The derivation of Feldt's coefficient requires congeneric part-tests and an assumption that error variances follow dictates of CTT, meaning the error variance for each part-test equals l_i times the error variance for the total test—this model is often referred to as *classical congeneric*. The formula for Feldt's coefficient can also be obtained by using the effective test lengths, $l_i = \sigma(X_i, X) / \sigma^2(X)$, in Equation 13.

Other congeneric-model coefficients that are not discussed here in detail include Kristof's (1974) coefficient for tests that can be divided into three congeneric part-tests with unknown lengths. Additionally, Gilmer and Feldt (1983) presented a more general formula applicable to tests with more than three congeneric parts with unknown lengths—it is referred to as the Feldt–Gilmer coefficient in Feldt and Brennan (1989). A maximum likelihood approach is also available, as described by Jöreskog (1971), through the use of a computer program LISREL (Jöreskog & Sörbom, 2018).

The internal consistency coefficients discussed thus far are suitable for tests consisting of items or part-tests that largely conform to the requirements of the essential tau-equivalence or congeneric models. Consider, however, a test composed of multiple content categories (or clusters), each with a distinct set of items. While items within each content category may satisfy the assumptions of

the essential tau-equivalence or congeneric model, applying these assumptions to items across content categories may lead to violations. This is because some unique true score variance may be associated with each content category. The approach discussed in the next section employs a general framework for linear composites to address situations where parallel-form assumptions hold reasonably well within clusters but not across them.

Reliability of Linear Composites

A linear composite is defined as a weighted sum of scores from multiple constituent components: $Z = \sum_{m=1}^M w_m X_m$, where w_m is the weight associated with component m and M is the number of components. In this definition, weights are unconstrained—they can be positive, negative, or even zero. The general framework of linear composites considered in this section does not require that the constituent components in a composite are parallel in the classical sense of parallelism. In effect, a composite can be derived from any types of components. Consider the following examples: a gain score, which is the difference between pretest and posttest scores; battery composites, which are usually formed by averaging multiple subtest scores (e.g., English, math, reading, and science); predicted scores based on a multiple linear regression; and total test scores obtained by summing scores from several distinct item types.

Conceptually, for a linear composite, each instance of a replication would involve the same set of components with items within components that vary. Each individual component, X_m , may consist of a set of items that conform, more or less, to one of the CTT definitions of parallelism and is allowed to have its own true, error, and observed score variances and, consequently, reliability. If independent errors across components are assumed, the error variance for composite scores is simply a weighted sum of all component error variances: $\sigma^2(E_Z) = \sum w_m^2 \sigma^2(E_{X_m})$. This error variance for the composite is incorporated in the error-variance definition of reliability to yield a general-purpose reliability coefficient for a linear composite:

$$R(Z)_{EV} = 1 - \frac{\sigma^2(E_Z)}{\sigma^2(Z)} = 1 - \frac{\sum w_m^2 \sigma^2(E_{X_m})}{\sigma^2(Z)} = 1 - \frac{\sum w_m^2 \sigma^2(X_m)[1 - R(X_m)]}{\sigma^2(Z)}, \quad (15)$$

where $R(X_m)$ is the reliability coefficient for component m . Equation 15 is not only convenient to use as long as a reliability estimate is available for each component, but also serves as a basis for deriving a reliability coefficient for many different types of linear composites. One such example is discussed next.

Stratified coefficient alpha (Rajaratnam et al., 1965) is generally considered more appropriate than the unstratified version, ${}_aR(X)$, when a test is organized according to a table of content (or other type) specifications. The assumption of essential tau equivalence may be sensible within content categories (called “strata” more generally), but may be questionable across them because there is presumably some unique true

score variance associated with each category. For estimation purposes, stratum-level scores are computed even though they may never be reported. Then, as a special case of Equation 15, stratified coefficient alpha is given by

$${}_{\text{strat } \alpha} R(Z) = 1 - \frac{\sum \sigma^2(X_m)[1 - {}_{\alpha} R(X_m)]}{\sigma^2(Z)}. \quad (16)$$

The numerator in Equation 16 is the error variance in ${}_{\text{strat } \alpha} R(Z)$, which is the sum of error variances in the regular coefficient alpha associated with the strata. Occasionally, unstratified ${}_{\alpha} R(X)$ is applied to stratified data without explicitly taking into account stratification. Doing so will likely yield an underestimate of reliability, although the extent of the problem depends on the nature of strata. Stratification by content or item types often makes meaningful differences.

The formula presented in Equation 15 is “general” in the sense that it does not specify which reliability coefficient should be used for the constituent components. If ${}_{\alpha} R(X)$ is used for all components, the result is stratified alpha. In a similar vein, a stratified version of Feldt’s coefficient may be obtained if one is willing to assume a congeneric model for the components. However, as W. Lee et al. (2025) demonstrated, stratification makes little difference between the stratified and unstratified Feldt’s coefficients. As a final note, nothing in theory excludes the possibility of using different reliability coefficients for various components, although it is rarely done in practice.

RELIABILITY IN GT

It is undoubtedly true that the very simple model of CTT provides an accessible framework for addressing various measurement problems. However, a principal limitation of the simple model is the presence of only one error term (E), where all sources of measurement error are clumped together and cannot be disentangled. As an extension and liberalization of CTT (Brennan, 2001c; Cronbach et al., 1972; Feldt & Brennan, 1989), GT permits the explicit modeling of multiple sources of random error. This feature makes GT a powerful and flexible tool for studying reliability and related issues. The multifacet aspect of GT is achieved through the application of analysis of variance (ANOVA) procedures. Other important characteristics that distinguish GT from CTT include the following: (a) GT incorporates a conceptual framework that differentiates between the concepts of *universes of admissible observations* (and associated G studies) and *universes of generalization* (and associated D studies); (b) in GT, a single notion of *randomly parallel* forms replaces the multiple definitions of parallelism in CTT; (c) GT introduces a clear distinction between two different types of error—absolute error and relative error; and (d) multivariate GT provides an even broader framework for

mixed models, including both random and fixed facets, with multiple universes of generalization.

The first comprehensive exploration of GT was presented in the book titled *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* by Cronbach et al. (1972). However, the application of ANOVA to measurement problems predates the formalization of GT by Cronbach's team. Earlier works by researchers such as Burt (1936, 1955), Ebel (1951), Hoyt (1941), and Lindquist (1953, chap. 16) had already introduced the idea. A more accessible exposition of the theory is provided by Shavelson and Webb (1991). Brennan's (2001c) book, titled *Generalizability Theory*, has been widely cited, serving as both a textbook and a resource for students and practitioners. For a comprehensive review of the history of GT up to the present, readers are directed to Brennan (2022) and the references therein.

We shall now delve into some fundamental concepts, definitions, and assumptions, primarily focusing on univariate GT. Following this, there will be a brief overview of multivariate GT, and the section will conclude with an introduction to extended multivariate GT. While a full treatment of GT is beyond the scope of this section, we will use a few hypothetical scenarios to elucidate some crucial concepts and methods.

Univariate GT

Univariate GT provides a framework to model and estimate distinct sources of random error separately. To draw a comparison with CTT, a tautological model for univariate GT can be expressed as

$$X_{pf} = \pi_p + (E_{pf1} + E_{pf2} + \cdots + E_{pfH}), \quad (17)$$

where π_p is the universe score for person p in the mean-score metric, which is analogous to the true score in CTT, and H is the number of sources of random error associated with "facets" (e.g., items, raters). The H error terms are closely tied to the investigator's intended universe of generalization (defined later), where replications ($f = 1, \dots, \infty$) are considered as comprising randomly parallel forms that differ with respect to the conditions of each facet. In this model, what constitutes error is a matter of definition and is emphatically under the control of the investigator.

Univariate GT is largely a two-step enterprise. The first step revolves around the notion of universes of admissible observations and generalizability studies (G studies), while the subsequent step considers the notion of universes of generalization and decision studies (D studies). These conceptual issues tend to be more challenging to grasp than the statistical issues concerning estimation of variance components and reliability statistics.

Universes of Admissible Observations and G Studies

Designing a measurement procedure begins with specifying a set of measurement conditions, known in GT as *facets*. To illustrate, consider a practical example of an English writing assessment. Among many possible considerations, one can identify facets of interest, such as writing prompts (denoted i) and raters (denoted r). Note that the investigator has full control over what and how many facets to consider given the specific purpose of the assessment. Both the numbers of potential writing prompts and the raters may be assumed to be indefinitely large, approaching infinity, at least theoretically. In this scenario, both facets are infinite in the *universe of admissible observations* (UAO), and the corresponding statistical model is called a random effects model, which is the primary focus of discussion in this section.³ Furthermore, if any prompt in the universe could be evaluated by any rater in the universe, the two facets in the UAO are crossed, denoted $i \times r$. Now suppose there is a target *population* of persons (p) for whom the writing assessment is intended. In this context, persons are the *objects of measurement* about whom decisions are drawn. If any prompt and any rater in their universes can be associated with any person in the population, it is symbolized as $p \times i \times r$, which characterizes any observable data for the population and universes.

A linear model associated with the $p \times i \times r$ structure that represents any observed score for a single person on a single prompt evaluated by a single rater is expressed as

$$X_{pir} = \mu + \nu_p + \nu_i + \nu_r + \nu_{pi} + \nu_{pr} + \nu_{ir} + \nu_{pir}, \quad (18)$$

where μ is the grand mean in the population and universes and the ν terms are mutually uncorrelated *effects*. The term ν_{pir} in Equation 18 represents a three-way interaction combined with all other residual effects. It follows that the variance of observed scores for a *single* condition of each facet can be decomposed into seven uncorrelated *variance components*:

$$\sigma^2(X_{pir}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(r) + \sigma^2(pi) + \sigma^2(pr) + \sigma^2(ir) + \sigma^2(pir). \quad (19)$$

These variance components play a central role in all subsequent analyses. A study is conducted to collect data for the purpose of estimating these variance components for the UAO. Such a study is called a generalizability study, or more succinctly, a G study. One possible data collection design would involve a sample of n_p persons, each of whom responds to the same set of n_i prompts that are evaluated by the same set of n_r raters. This is called a $p \times i \times r$ G-study design, which is consistent with the structure of the UAO.

However, the UAO does not dictate a particular G-study design for collecting observed data. For example, if each of n_r raters provides ratings on the responses to two or more nonoverlapping prompts, this is a design symbolized as $p \times (i:r)$,

where prompts are “nested” within raters. While there is no definitive answer to which G-study design should be used, it is generally preferred to use a crossed design (i.e., $p \times i \times r$) when the UAO is crossed. Using a crossed design provides the benefit that all seven variance components in Equation 19 can be independently estimated, offering great flexibility in obtaining results for other nested designs as well as the crossed design. By contrast, using a nested G-study design, some variance components in Equation 19 are not separately estimable (i.e., confounded), and results can only be obtained for certain nested designs but not for the crossed design.

Based on the G-study data, estimates of variance components can be obtained using various methods, including the ANOVA procedure. This procedure essentially involves the explicit use of the expected mean-square equations for a given G-study design, setting them equal to the weighted sum of contributing variance components. The expected mean squares are substituted with observed mean squares in the equations, and they are then solved for each of the variance components in terms of the mean squares. Once estimates of the variance components are obtained, they can be used to estimate error variances and reliability-like coefficients in subsequent decision studies, as discussed next.

Universes of Generalization and D Studies

As indicated in Equation 19, the G-study variance components for the UAO pertain to scores on *single* conditions of each facet. These estimated G-study variance components can be used in designing efficient measurement procedures for operational purposes through various decision (D) studies. Conceivably, an operational measurement procedure might encompass a set of measurement conditions (e.g., four prompts and two raters) from the UAO, and decisions about objects of measurement (persons) will be based on their *mean* (i.e., average) scores over those measurement conditions. Moreover, there exists a large (presumably infinite) number of similar forms that can be selected from the infinite UAO, each form containing different samples of measurement conditions (i.e., different sets of four prompts and two raters). Such forms are referred to as *randomly parallel forms*, which embody the notion of replications in GT. Unlike the definitions of parallelism in CTT, no constraints such as equal means and variances are imposed on randomly parallel forms, which are essentially indistinguishable from one another according to the principle of random sampling.

The concept of replications tied to randomly parallel forms is intricately linked to the notion of *universe of generalization* (UG). Indeed, a UG constitutes the universe of randomly parallel forms to which the investigator intends to generalize the results based on a particular form. In essence, a UAO is a universe for single observations, whereas a UG is a test-level universe focusing on mean scores over sets of measurement conditions. In a similar vein, the G-study variance components are derived from single-observation data and are subsequently employed to obtain the D-study

variance components for mean scores over sets of conditions. The design structure of the D study may or may not mirror that of the G study. Furthermore, it is entirely legitimate for a UG to contain either all or a subset of conditions present in the UAO. These conceptual considerations are illustrated using the same example mentioned earlier.

Consider a scenario where the UAO is infinite with the $p \times i \times r$ structure, and G-study variance components have been estimated using the $p \times i \times r$ design with n_i prompts and n_r raters. Now, consider the following three different scenarios for D studies:

1. Same design and same universe as G study
 - Design: $p \times I \times R$
 - UG: both I and R are random (i.e., contains all the conditions in the UAO)
2. Same design but different universe
 - Design: $p \times I \times R$
 - UG: I is random, but R is a fixed facet (i.e., interest focuses on generalizing results over prompts only, but not raters)
3. Different design with same universe
 - Design: $p \times (I : R)$
 - UG: both I and R are random

Note that uppercase letters represent facets in D studies (i.e., I and R) to emphasize the use of *mean* scores over conditions, as opposed to the single prompt-rater scores in the G study. In addition, the D-study sample sizes for the prompts and raters can be user defined and may be the same as or different from the G-study sample sizes, denoted n'_i and n'_r .

It is particularly important to note that the specification of a UG and characterization of a D-study design bear resemblance to the *conceptualization-estimation* framework for reliability discussed in the section “Terms and Background.” To define a UG, an investigator needs to articulate or conceptualize what constitutes a replication of a measurement procedure, predominantly in terms of which facets are considered random and which are fixed. The features of a D-study design pertain to estimation of reliability statistics.

The linear model for the first D-study scenario (i.e., $p \times I \times R$ with both I and R random) representing a person's observed mean score over n'_i prompts and n'_r raters can be expressed as

$$X_{pIR} = \bar{X}_p = \mu + v_p + v_I + v_R + v_{pI} + v_{pR} + v_{IR} + v_{pIR}, \quad (20)$$

which is analogous to Equation 18 except for the use of uppercase subscripts, I and R . The D-study variance components for the score effects in Equation 20 can be obtained by dividing the corresponding G-study variance components by the user-defined D-study sample sizes. In this example, $\sigma^2(I) = \sigma^2(i)/n'_i$, $\sigma^2(R) = \sigma^2(r)/n'_r$,

$$\sigma^2(pI) = \sigma^2(pi)/n'_i, \quad \sigma^2(pR) = \sigma^2(pr)/n'_r, \quad \sigma^2(IR) = \sigma^2(ir)/n'_i n'_r, \quad \text{and} \\ \sigma^2(pIR) = \sigma^2(pir)/n'_i n'_r.$$

As noted earlier, GT employs the expected-value perspective on true score, which it refers to as *universe score*. That is, universe score is defined as the expected value of observed mean scores over all randomly parallel forms in the UG:

$$\pi_p \equiv \mu_p \equiv \mathcal{E}_T \mathcal{E}_R X_{pIR}. \quad (21)$$

The variance of universe scores over all persons in the population is defined as $\sigma^2(\pi) = \mathcal{E}_p (\mu_p - \mu)^2$. For the $p \times I \times R$ design with random I and R , the universe score variance is simply $\sigma^2(\pi) = \sigma^2(p)$.

The second scenario differs from the first in that its UG is more narrowly defined than the UAO by fixing the rater facet. In this case, every randomly parallel form consists of the exact same set of raters, eliminating generalization over raters. The net effect is that the variance component for the person-by-rater interaction effect contributes to universe score variance; specifically, $\sigma^2(\pi) = \sigma^2(p) + \sigma^2(pR)$.

For the third scenario with a $p \times (I : R)$ D-study design, the linear model is

$$X_{pIR} = \bar{X}_p = \mu + v_p + v_R + v_{I:R} + v_{pR} + v_{pI:R}. \quad (22)$$

This linear model has fewer score effects (and variance components) compared to the $p \times I \times R$ design shown in Equation 20. If the G-study variance components for the fully crossed $p \times i \times r$ design are available, another G study is not required to obtain the D-study variance components for the nested design. Instead, a nested effect can be expressed in terms of confounded effects from the crossed design; that is, $\sigma^2(i : r) = \sigma^2(i) + \sigma^2(ir)$ and $\sigma^2(pi : r) = \sigma^2(pi) + \sigma^2(pir)$. It follows that $\sigma^2(I : R) = [\sigma^2(i) + \sigma^2(ir)]/n'_i n'_r$ and $\sigma^2(pI : R) = [\sigma^2(pi) + \sigma^2(pir)]/n'_i n'_r$ — the other three variance components are the same for both designs. Although the design for the third scenario differs from the first one, universe score variance remains unchanged (i.e., $\sigma^2(p)$) because the UG is the same.

Error Variances and Coefficients

Reliability coefficients in CTT are fundamentally correlation-based, directly or indirectly focusing on the rank ordering of test takers. Consequently, error variance associated with CTT reliability coefficients is interpreted in relation to the performance of other persons in the group, making it inherently *relative*. By contrast, GT distinguishes between two types of error: absolute error and relative error, each with different interpretations and uses. Absolute error for a person is the difference between their observed mean score and the universe score:

$$\Delta_p \equiv \bar{X}_p - \mu_p. \quad (23)$$

The variance of Δ_p over persons is referred to as *absolute error variance*, denoted $\sigma^2(\Delta)$, which is more suitable for criterion-referenced interpretations of scores. On the contrary, relative error, akin to CTT, is the difference between observed deviation scores and universe deviation scores:

$$\delta_p \equiv (\bar{X}_p - \mathcal{E}_p \bar{X}_p) - (\mu_p - \mu). \quad (24)$$

Relative error variance, $\sigma^2(\delta)$, is the variance of Equation 24 over persons.

Each type of error variance has an associated reliability-like coefficient. A generalizability coefficient involves relative error variance as its error term, defined as

$$\mathcal{E}\rho^2 = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\delta)}. \quad (25)$$

A different coefficient that involves absolute error variance is called a dependability coefficient, given by

$$\Phi = \frac{\sigma^2(\pi)}{\sigma^2(\pi) + \sigma^2(\Delta)}. \quad (26)$$

Results for error variances and coefficients depend on a specified UG, a D-study design, and corresponding D-study variance components. Relative error variance, $\sigma^2(\delta)$, and a generalizability coefficient, $\mathcal{E}\rho^2$, are analogous to error variance and a reliability coefficient, respectively, in CTT. For a simple $p \times I$ design, $\mathcal{E}\rho^2$ is equal to coefficient alpha (or KR20 for binary items) if $n'_i = n_i$; however, this equality does not extend to other UGs and/or D-study designs.

It can be shown that absolute error variance is the sum of all variance components except $\sigma^2(p)$ for a random model with a crossed design. In the first D-study scenario with the $p \times I \times R$ design (both I and R random),

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(R) + \sigma^2(pI) + \sigma^2(pR) + \sigma^2(IR) + \sigma^2(pIR).$$

Relative error variance, by contrast, contains all variance components for interaction effects that involve p :

$$\sigma^2(\delta) = \sigma^2(pI) + \sigma^2(pR) + \sigma^2(pIR).$$

Clearly, $\sigma^2(\delta)$ is smaller than $\sigma^2(\Delta)$, which, in turn, leads to $\mathcal{E}\rho^2$ that is larger than Φ .

In the second scenario with R fixed, $\sigma^2(pR)$ contributes to universe score variance, resulting in

$$\sigma^2(\Delta) = \sigma^2(I) + \sigma^2(pI) + \sigma^2(IR) + \sigma^2(pIR)$$

$$\sigma^2(\delta) = \sigma^2(pI) + \sigma^2(pIR).$$

Compared to the first scenario, this leads to larger universe score variance, smaller error variances, and larger coefficients. This outcome is a consequence of the reduced error in a restricted UG that limits the generalization of results.

For the third scenario, involving an infinite UG with a $p \times (I : R)$ design,

$$\begin{aligned}\sigma^2(\Delta) &= \sigma^2(R) + \sigma^2(I : R) + \sigma^2(pR) + \sigma^2(pI : R) \\ \sigma^2(\delta) &= \sigma^2(pR) + \sigma^2(pI : R).\end{aligned}$$

Results for this design are likely to differ from those of the first scenario with a fully crossed design, even if both share the same UG. This underscores the dependence of D-study results on both UG characteristics and D-study design structure.

The hypothetical scenarios with two-facet designs served as a vehicle to introduce key concepts, terminology, and statistical formulas in univariate GT. They also provided insight into the considerations necessary for establishing a well-defined measurement procedure. Perhaps the most critical consideration, after identifying all pertinent sources of error (i.e., facets), is determining which facets are random and which are fixed in the intended UG. Other factors demanding attention include the D-study design structure and D-study sample sizes. All these considerations affect D-study results, one way or another.

Multivariate GT

An early exploration of multivariate GT was presented by Rajaratnam et al. (1965), particularly in the context of introducing stratified alpha as a tool for estimating reliability for a composite. Considered as a precursor to multivariate GT, stratified alpha set the stage for later developments. Later, Cronbach et al. (1972) offered the first integrated treatment of multivariate GT, a framework further expanded by Brennan (2001c).

Expanding on the univariate model depicted in Equation 17, multivariate GT may be represented as

$$X_{pf} = (\pi_{p1} + \pi_{p2} + \cdots + \pi_{pM}) + (E_{pf1} + E_{pf2} + \cdots + E_{pfH}), \quad (27)$$

in which each person possesses M universe scores, each associated with one of M fixed conditions of measurement; and, as previously defined, the number of sources of error (i.e., facets), denoted by H , is typically the same across all M levels of a fixed facet—more intricate multivariate designs are considered in the section “Extended Multivariate GT.” While many fundamental principles of univariate GT still apply to multivariate GT, a key distinction lies in the presence of “multiple” univariate random effects models, each corresponding to each level of the fixed facet.

To apply multivariate GT, there must be at least one fixed facet and at least one random facet in the UAO and G-study data collection design. For illustration, let us consider a simple example based on the so-called table of specifications model

(Jarjoura & Brennan, 1982, 1983). In this model, each item in the UAO is associated with one of several specific content categories, making items (i) random and content categories (c) fixed. Suppose a G study is conducted to gather response data from a group of test takers (p) who respond to all items nested within each content category. This multivariate design is designated $p^{\bullet} \times i^{\circ}$, where the filled circle superscript indicates that the facet is crossed with the levels of the fixed facet, c , and the empty circle superscript signifies the facet nested within c .

For simplicity, suppose there are two content categories (i.e., $n_c=2$). Under the multivariate $p^{\bullet} \times i^{\circ}$ design, covariances exist for the object of measurement (persons) since the same persons respond to all items in both content categories. However, there are no covariances for items that are nested within c . The variance and covariance components for p , i , and pi are displayed in the following $n_c \times n_c$ matrices:

$$\Sigma_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}, \Sigma_i = \begin{bmatrix} \sigma_1^2(i) & \\ & \sigma_2^2(i) \end{bmatrix}, \text{ and } \Sigma_{pi} = \begin{bmatrix} \sigma_1^2(pi) & \\ & \sigma_2^2(pi) \end{bmatrix},$$

where the subscripts 1 and 2 denote the two content categories. Note that both Σ_i and Σ_{pi} are diagonal matrices with zero covariances. Since there is a univariate random $p \times i$ design for each content level, the variances reported in the first and second columns in the matrices represent the variance components for a $p \times i$ design based solely on data from the first and second content categories, respectively. Hence, the conventional ANOVA procedure can be used to estimate univariate variance components for each content category. The estimate of the covariance component in Σ_p , $\sigma_{12}(p)$, is simply the covariance between persons' observed scores on the two categories. For more complex designs, interested readers can refer to Brennan (2001c, chaps. 9 and 11) for procedures on estimating covariance components.

Now, consider a subsequent D study that involves n'_{i1} items for Category 1 and n'_{i2} items for Category 2 with the same design structure as the G study. In this case, the two diagonal elements in the variance-covariance component matrix for I , Σ_I , are $\sigma_1^2(I) = \sigma_1^2(i)/n'_{i1}$ and $\sigma_2^2(I) = \sigma_2^2(i)/n'_{i2}$. Likewise, Σ_{pI} contains the following diagonal elements: $\sigma_1^2(pI) = \sigma_1^2(pi)/n'_{i1}$ and $\sigma_2^2(pI) = \sigma_2^2(pi)/n'_{i2}$. It follows that the matrices for universe score, relative error, and absolute error, respectively, are $\Sigma_{\pi} = \Sigma_p$; $\Sigma_{\delta} = \Sigma_{pi}$; and Σ_{Δ} with diagonal elements of $\sigma_1^2(\Delta) = \sigma_1^2(I) + \sigma_1^2(pI)$ and $\sigma_2^2(\Delta) = \sigma_2^2(I) + \sigma_2^2(pI)$. Note that for this design, Σ_{Δ} and Σ_{δ} are diagonal matrices.

In practical applications, decisions about test takers are often based on their composite scores over all levels of content categories. A composite based on two content categories can be expressed as a weighted sum of two mean scores: $\bar{Z} = w_1\bar{X}_1 + w_2\bar{X}_2$, where the bar above Z denotes the mean-score metric, and weights are typically

proportional to the number of items in each category (i.e., $w_1 = n'_{i1} / (n'_{i1} + n'_{i2})$ and $w_2 = n'_{i2} / (n'_{i1} + n'_{i2})$). Similarly, the composite universe score for a person is defined as $\pi_{p\bar{z}} = w_1\pi_{p1} + w_2\pi_{p2}$. Since $\Sigma_{\pi} = \Sigma_p$, composite universe score variance $\sigma_{\bar{z}}^2(\pi)$ is the sum of all of the elements in Σ_p with associated weights: $\sigma_{\bar{z}}^2(\pi) = w_1^2\sigma_1^2(p) + w_2^2\sigma_2^2(p) + 2w_1w_2\sigma_{12}(p)$. Relative error variance and absolute error variance for the composite scores, respectively, are weighted sums of diagonal elements of their corresponding matrices, as follows: $\sigma_{\bar{z}}^2(\delta) = w_1^2\sigma_1^2(\delta) + w_2^2\sigma_2^2(\delta)$ and $\sigma_{\bar{z}}^2(\Delta) = w_1^2\sigma_1^2(\Delta) + w_2^2\sigma_2^2(\Delta)$. These error variances, when combined with composite universe score variance, provide reliability-like coefficients for the composite scores. A multivariate generalizability coefficient and a multivariate dependability coefficient for the composite are defined similarly to Equations 25 and 26, respectively.

The typical computational procedures for multivariate GT designs have been demonstrated using a simple $p^{\bullet} \times i^{\circ}$ design so far. Applying these procedures to multifacet situations and/or different design structures is straightforward as long as the multivariate design is properly identified. For example, consider a scenario with two types (c) of short free-response items (i) that are evaluated by raters (r), where each rater scores test takers' (p) responses to a *different* subset of items for both types. This constitutes a multivariate $p^{\bullet} \times (i^{\circ} : r^{\bullet})$ design with c serving as a multivariate fixed facet. Items are nested within both r and c , and for each level of c , there is a univariate random $p \times (i : r)$ design. Some effects (p , r , and pr) will have full matrices with nonzero covariance terms, while other effects ($i : r$ and $pi : r$) will have diagonal matrices with zero covariances. D-study variance and covariance components, along with various D-study statistics, can then be computed using the procedures delineated in the preceding discussion of the simplest design.

The procedures discussed thus far are generally sufficient for most applications with a "single" multivariate design where the same univariate design structure applies to all levels of the fixed facet. Nonetheless, more complex designs can arise in practical situations, necessitating special treatment, as discussed in the following section.

Extended Multivariate GT

The extant literature on GT offers limited guidance on handling a mixture of different design structures in a multivariate design. One such example is a mixed-format exam that contains both multiple-choice (MC) items and free-response (FR) items involving raters for scoring. Here, the item type represents a multivariate fixed facet with two levels, MC and FR. Moses and Kim (2015) were the first to consider combining two distinct designs in multivariate GT, and Brennan et al. (2022) later provided a more extensive treatment, referring to it as *extended* multivariate GT.

The model representation for conventional multivariate GT shown in Equation 27 may be further expanded to characterize extended multivariate GT. Specifically,

$$X_{pf} = [\pi_{p1} + (E_{pf1} + \cdots + E_{pfH_1})] + \cdots + [\pi_{pM} + (E_{pf1} + \cdots + E_{pfH_M})]. \quad (28)$$

This model suggests that there are M distinct levels of a fixed facet, each associated with its own sources of error (i.e., facets) signified by different subscripts to H , H_1 through H_M . Since this model has not been fully exploited in the literature, one of the examples reported by Brennan et al. (2022) is briefly introduced here.

To continue with the example of a mixed-format exam, the MC section involves items (i) as the primary measurement condition, while the FR section involves both items (i) and raters (r). For notational simplicity, the symbol i is used to represent both MC and FR items—however, it does not imply interchangeability between the two item types. Representing the combination of two universes, the UAO structure can be symbolized as $\{p^\bullet \times i^\circ\}\{p^\bullet \times i^\circ \times r^\circ\}$, where persons are crossed with both universes, as indicated by the closed circle superscript. Brennan et al. (2022) concisely represented this design as $p^\bullet \times [i^\circ \cup (i^\circ \times r^\circ)]$, where the symbol $\cup$ denotes the union of the MC and FR sections, and the first and second i° before and after $\cup$ are associated with the MC and FR sections, respectively.

Suppose G-study data were collected for the $p^\bullet \times [i^\circ \cup (i^\circ \times r^\circ)]$ design. The conventional univariate methods can be applied to estimate G-study variance components separately for the MC and FR sections based on a univariate $p \times i$ design for the MC section and a univariate $p \times i \times r$ design for the FR section. For the full multivariate design, the covariance component for p needs to be estimated. Given that p is the only linked facet in this design, an unbiased estimate of the covariance is simply the observed covariance between MC and FR scores.

The left side of Table 5.1 presents the G-study variance and covariance components for this multivariate design. Each matrix reports variance components for MC and FR (indicated by the subscripts 1 and 2) in the first and second columns, respectively. Notably, the MC section, analyzed with a $p \times i$ design, involves three relevant variance components (p , i , and pi). In contrast, the FR section, analyzed with a $p \times i \times r$ design, includes seven variance components (p , i , r , pi , pr , ir , and pir). Each of the seven matrices contains variance components corresponding to one of the seven effects for FR. However, only three matrices (Σ_p , Σ_i , and Σ_{pi}) report variance components for MC. This is because the other four matrices are not relevant to MC, as indicated by “NA.” Covariance components are present in Σ_p only.

On the right side of Table 5.1, the D-study variance and covariance components are presented. These components are obtained by dividing their corresponding G-study components by the user-defined D-study sample sizes, n'_{i1} , n'_{i2} , and n'_r . Composite universe score variance is calculated as the weighted sum of all the elements in Σ_p : $\sigma^2_{\bar{z}}(\pi) = w_1^2 \sigma_1^2(p) + w_2^2 \sigma_2^2(p) + 2w_1 w_2 \sigma_{12}(p)$. Since MC and FR involve different sources of error, both relative error variance and absolute error variance for the

Table 5.1 Variance and Covariance Components for the Extended Multivariate Design Example

G study $p^\bullet \times [i^\circ \cup (i^\circ \times r^\circ)]$		D study $p^\bullet \times [I^\circ \cup (I^\circ \times R^\circ)]$	
MC	FR	MC	FR
$\Sigma_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}$		$\Sigma_p = \begin{bmatrix} \sigma_1^2(p) & \sigma_{12}(p) \\ \sigma_{12}(p) & \sigma_2^2(p) \end{bmatrix}$	
$\Sigma_i = \begin{bmatrix} \sigma_1^2(i) & \\ & \sigma_2^2(i) \end{bmatrix}$		$\Sigma_I = \begin{bmatrix} \frac{\sigma_1^2(i)}{n'_{i1}} & \\ & \frac{\sigma_2^2(i)}{n'_{i2}} \end{bmatrix}$	
$\Sigma_r = \begin{bmatrix} \text{NA} & \\ & \sigma_2^2(r) \end{bmatrix}$		$\Sigma_R = \begin{bmatrix} \text{NA} & \\ & \frac{\sigma_2^2(r)}{n'_r} \end{bmatrix}$	
$\Sigma_{pi} = \begin{bmatrix} \sigma_1^2(pi) & \\ & \sigma_2^2(pi) \end{bmatrix}$		$\Sigma_{pI} = \begin{bmatrix} \frac{\sigma_1^2(pi)}{n'_{i1}} & \\ & \frac{\sigma_2^2(pi)}{n'_{i2}} \end{bmatrix}$	
$\Sigma_{pr} = \begin{bmatrix} \text{NA} & \\ & \sigma_2^2(pr) \end{bmatrix}$		$\Sigma_{pR} = \begin{bmatrix} \text{NA} & \\ & \frac{\sigma_2^2(pr)}{n'_r} \end{bmatrix}$	
$\Sigma_{ir} = \begin{bmatrix} \text{NA} & \\ & \sigma_2^2(ir) \end{bmatrix}$		$\Sigma_{IR} = \begin{bmatrix} \text{NA} & \\ & \frac{\sigma_2^2(ir)}{n'_{i2} n'_r} \end{bmatrix}$	

$\Sigma_{pir} = \begin{bmatrix} NA & \\ & \sigma_2^2(pir) \end{bmatrix}$	$\Sigma_{pIR} = \begin{bmatrix} NA & \\ & \frac{\sigma_2^2(pir)}{n'_{i2} n'_r} \end{bmatrix}$

Note. D study = decision study; FR = free response; G study = generalizability study; MC = multiple choice.

composite involve different variance and covariance components for the two sections, as follows:

$$\sigma_{\bar{Z}}^2(\delta) = w_1^2[\sigma_1^2(pI)] + w_2^2[\sigma_2^2(pI) + \sigma_2^2(pR) + \sigma_2^2(pIR)], \quad (29)$$

and

$$\begin{aligned} \sigma_{\bar{Z}}^2(\Delta) = & w_1^2[\sigma_1^2(I) + \sigma_1^2(pI)] + w_2^2[\sigma_2^2(I) + \sigma_2^2(R) + \sigma_2^2(IR) \\ & + \sigma_2^2(pI) + \sigma_2^2(pR) + \sigma_2^2(pIR)]. \end{aligned} \quad (30)$$

Finally, the reliability-like coefficients are obtained using these quantities for the composite in Equations 25 and 26.

The preceding discussion highlights that extended multivariate GT offers a useful framework for dealing with multivariate designs that are more complex than those commonly discussed in the current literature. However, the model needs to be studied further through more diversified applications of the model with various types of complex designs in real-world testing.

RELIABILITY IN IRT

In IRT, a test taker's response to an item is modeled as a function of an underlying proficiency variable denoted θ and item parameters. The proficiency for a test taker can be estimated using either the maximum likelihood (ML) or Bayesian estimation methods. An estimator, denoted $\hat{\theta}$, serves as the observed score in IRT. In this context, we primarily consider the ML and Bayesian expected a posteriori (EAP) estimators. Another Bayesian estimator, maximum a posteriori (MAP), is briefly discussed because it shares the same underlying framework with the EAP estimator.

The main focus of this section is to present various approaches to quantifying reliability within the framework of IRT. The mathematical formulations presented here are designed to be applicable and generalizable to many different IRT models, without making specific distinctions among those models. Throughout this section, unidimensional dichotomous IRT models are assumed, except in the last subsection where some procedures are discussed for a few specific multidimensional IRT (MIRT) models.

Results are presented for several score metrics, including ML estimates, EAP estimates, summed raw scores, transformed scale scores from summed raw scores, composite scores, and composite scale scores. The two classes of estimation methods (ML and EAP) are discussed separately because of their fundamental inconsistencies in terms of the model framework and assumptions.

Under the ML framework, the data at hand are viewed as a random sample from a distribution with known θ and item parameters. For a test taker with θ , the observed data over replications are conceptually obtained by administering test forms with the exact same item parameters repeatedly. These forms, with the same item parameters, are often called *strictly parallel* forms, or it can be said that the form is effectively fixed (W. Lee et al., 2000). Therefore, the source of error causing the variability in estimates over replications is best described as resulting from each item being sampled from a large pool of items with identical parameters.

Reliability statistics under the ML framework are conceptualized and computed by focusing on the distribution of proficiency estimates conditional on the parameter $f(\hat{\theta}|\theta)$. The standard deviation of $\hat{\theta}$ given θ is termed the conditional standard error of estimation (SEE). In IRT, the SEE plays a role similar to that of the SEM in CTT. However, these terms are often used interchangeably in the context of estimating reliability in IRT (e.g., B. F. Green et al., 1984; Lord, 1983). For the purposes of this chapter, no distinction is made between the two terms.

In contrast, the Bayesian framework is primarily concerned with obtaining the posterior distribution of parameters given observed data, $f(\theta|\mathbf{u})$, where parameters are considered random while data are fixed. Here, $\mathbf{u}$ is a vector of a test taker's item responses. The EAP estimate is the mean of the posterior distribution of θ for test takers with the same item responses. The standard deviation of the posterior distribution serves as the SEE for the EAP estimator. However, as demonstrated later, under this Bayesian framework, the concept of replications is obscure, at best, and thus the Bayes SEE is conceptually different from the SEE in the ML estimator and SEM in CTT.

Consider the following parameterization, similar to CTT:

$$\hat{\theta} = \theta + \varepsilon, \quad (31)$$

where a proficiency estimate $\hat{\theta}$ is decomposed into the latent proficiency θ and an error component, ε . For simplicity, subscripts indicating test takers and forms are omitted here. In contrast to CTT, where the true score for a test taker is equal to the expected value of observed scores, the proficiency estimators, both ML and EAP, provide biased estimates for θ (J. K. Kim & Nicewander, 1993; Lord, 1983; Warm, 1989). The ML estimator is unbiased only asymptotically as the test length approaches infinity. The bias in the ML estimates has significant consequences. Unlike the results in CTT, various expressions or definitions of the reliability coefficient for the ML estimator may not yield the same answer, a fact that is frequently overlooked.

Maximum Likelihood Proficiency Estimates

Let $\hat{\theta}$ denote the ML estimator of θ . As noted earlier, due to bias in $\hat{\theta}$, the expected value of $\hat{\theta}$ is not equal to θ ; namely, $\mathcal{E}(\hat{\theta}|\theta) \neq \theta$. The expected value of $\hat{\theta}$ is referred to here as *expected* proficiency, compared to *latent* proficiency θ . For $\hat{\theta}$, the following reliability coefficients are typically considered:

$$R(\hat{\theta})_{PF} \equiv \rho(\hat{\theta}, \hat{\theta}'); R(\hat{\theta})_{SC} \equiv \rho^2(\hat{\theta}, \theta); R(\hat{\theta})_{VR} \equiv \frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})};$$

$$R(\hat{\theta})_{EVR} \equiv \frac{\sigma^2[\mathcal{E}(\hat{\theta}|\theta)]}{\sigma^2(\hat{\theta})} = 1 - \frac{\mathcal{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})}; \text{ and } R(\hat{\theta})_{MVR} \equiv \frac{\sigma^2(\theta)}{\sigma^2(\theta) + \mathcal{E}\sigma^2(\hat{\theta}|\theta)}.$$

In CTT, similar expressions of these reliability coefficients are all identical when derived under the assumptions of classically parallel forms. However, the equality does not hold for $\hat{\theta}$. The first definition $R(\hat{\theta})_{PF}$ is the parallel-forms coefficient, $R(\hat{\theta})_{SC}$ is the squared-correlation coefficient, and the last three represent the variance-ratio coefficients (indicated by subscripts including VR). These coefficients will likely produce different results primarily because they use different definitions of true proficiency and its variance. Hereinafter, $\sigma^2(\theta)$ in $R(\hat{\theta})_{VR}$ and $R(\hat{\theta})_{MVR}$ is referred to as *latent proficiency variance*, while the numerator of the first formula in $R(\hat{\theta})_{EVR}$, $\sigma^2[\mathcal{E}(\hat{\theta}|\theta)]$, is referred to as *expected proficiency variance*. The subscript *EVR* appended to $R(\hat{\theta})_{EVR}$ signifies the use of expected proficiency variance. Likewise, $\mathcal{E}\sigma^2(\hat{\theta}|\theta)$ in $R(\hat{\theta})_{EVR}$ and $R(\hat{\theta})_{MVR}$ is referred to here as *expected error variance*, which differs from either the error variance $\sigma^2(\varepsilon)$ (see Equation 31) or another quantity, $\sigma^2(\hat{\theta}) - \sigma^2(\theta)$. Recall that $\mathcal{E}\sigma^2(\hat{\theta}|\theta)$ quantifies the variability of $\hat{\theta} - \mathcal{E}\hat{\theta}$; by contrast, $\sigma^2(\varepsilon)$ is concerned about $\hat{\theta} - \theta$. The last three variance-ratio coefficients differ in terms of which types of variances are used in the numerators and denominators. In effect, $R(\hat{\theta})_{VR}$ involves latent proficiency variance in conjunction with $\sigma^2(\hat{\theta}) - \sigma^2(\theta)$ as its error variance; $R(\hat{\theta})_{EVR}$ contains expected variances in both the numerator and denominator; and $R(\hat{\theta})_{MVR}$ exploits an interesting mix of both types (note the letter *M* in the subscript).

For ease of explanation, let's first delve into the three variance-ratio coefficients, followed by $R(\hat{\theta})_{PF}$ and $R(\hat{\theta})_{SC}$. In the literature, $R(\hat{\theta})_{EVR}$ is often called marginal reliability (B. F. Green et al., 1984). S. Kim (2012) labeled $R(\hat{\theta})_{EVR}$ as the (squared) correlation ratio for predicting θ from $\hat{\theta}$, which might also be called an intraclass correlation coefficient based on the ANOVA identity, given by

$$\sigma^2(\hat{\theta}) = \sigma^2[\mathcal{E}(\hat{\theta}|\theta)] + \mathcal{E}\sigma^2(\hat{\theta}|\theta). \quad (32)$$

The intraclass correlation coefficient, $R(\hat{\theta})_{EVR}$, becomes identical to the squared Pearson correlation coefficient (i.e., $R(\hat{\theta})_{SC}$) if the relationship between $\hat{\theta}$ and θ is linear; however, if the relationship is not linear, $R(\hat{\theta})_{EVR} \geq R(\hat{\theta})_{SC}$.

Since $\hat{\theta}$ is a biased estimator, $\mathcal{E}(\hat{\theta}|\theta) = \theta + BIAS$. Bias is not constant and can be either positive or negative depending on the location of a proficiency parameter along the θ scale. An immediate consequence of the bias in $\hat{\theta}$ is that $\sigma^2[\mathcal{E}(\hat{\theta}|\theta)] \neq \sigma^2(\theta)$. Rather,

$$\sigma^2[\mathcal{E}(\hat{\theta}|\theta)] = \sigma^2(\theta + BIAS) = \sigma^2(\theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS), \quad (33)$$

where the variance terms are always positive and the covariance between θ and $BIAS$ is usually positive although there is no theoretical justification or proof for that (Lord, 1983). Clearly, the expected proficiency variance, $\sigma^2[\mathcal{E}(\hat{\theta}|\theta)]$, is *not* equal to $\sigma^2(\theta)$ and the former usually is larger than the latter. From Equations 32 and 33, it follows that

$$\sigma^2(\hat{\theta}) = \sigma^2(\theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS) + \mathcal{E}\sigma^2(\hat{\theta}|\theta). \quad (34)$$

As $\sigma(\theta, BIAS)$ typically is positive, $\sigma^2(\hat{\theta}) > \sigma^2(\theta)$.

Coefficient $R(\hat{\theta})_{EVR}$ uses the expected proficiency variance in the numerator and the expected error variance in the denominator because $\sigma^2(\hat{\theta}) - \sigma^2[\mathcal{E}(\hat{\theta}|\theta)] = \mathcal{E}\sigma^2(\hat{\theta}|\theta)$. Conceptually, $BIAS = \mathcal{E}(\hat{\theta}|\theta) - \theta$ is a constant for a given test taker with θ (i.e., systematic error), and thus its variance (and covariance) should contribute to the true proficiency variance, but not to the error variance. Equations 33 and 34 clearly show that the expected proficiency variance, $\sigma^2[\mathcal{E}(\hat{\theta}|\theta)]$, in $R(\hat{\theta})_{EVR}$ incorporates two additional terms related to bias, in addition to $\sigma^2(\theta)$, whereas the expected error variance does not encompass them.

To estimate $R(\hat{\theta})_{EVR} = 1 - [\mathcal{E}\sigma^2(\hat{\theta}|\theta)/\sigma^2(\hat{\theta})]$, the conditional error variance, $\sigma^2(\hat{\theta}|\theta)$, is often obtained by taking the reciprocal of the test information function as $\sigma^2(\hat{\theta}|\theta) = 1 / I(\theta, \hat{\theta})$. Thus, $R(\hat{\theta})_{EVR}$ can also be expressed as

$$R(\hat{\theta})_{EVR} = 1 - \frac{\mathcal{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} = 1 - \frac{\mathcal{E}[1/I(\theta, \hat{\theta})]}{\sigma^2(\hat{\theta})}. \quad (35)$$

Taking the three-parameter logistic IRT model as an example, the test information on an n -item test is simply the sum of the item information functions:

$$I(\theta, \hat{\theta}) = \sum_{i=1}^n I_i(\theta, \hat{\theta}) = \sum_{i=1}^n \left[\frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)} \right], \text{ where } I_i(\theta, \hat{\theta}) (i = 1, \dots, n)$$

is the item information functions, $P_i(\theta) = c_i + (1 - c_i) / \{1 + \exp[-1.7a_i(\theta - b_i)]\}$, and $Q_i(\theta) = 1 - P_i(\theta)$. The expectation in Equation 35 typically is taken over a set of discrete θ quadrature points and associated weights assuming a standard normal distribution.

Coefficient $R(\hat{\theta})_{VR}$ differs from $R(\hat{\theta})_{EVR}$ in its use of latent proficiency variance and a distinct form of error variance expressed as $\sigma^2(\hat{\theta}) - \sigma^2(\theta)$:

$$\sigma^2(\hat{\theta}) - \sigma^2(\theta) = \mathcal{E}\sigma^2(\hat{\theta}|\theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS). \quad (36)$$

This “error variance” comprises the expected error variance plus two terms associated with systematic error, namely, bias. The bias-related terms, which contribute to the expected proficiency variance in $R(\hat{\theta})_{EVR}$, are integrated into the error variance in $R(\hat{\theta})_{VR}$, resulting in $R(\hat{\theta})_{EVR} \geq R(\hat{\theta})_{VR}$. This inclusion of bias terms in the error variance of $R(\hat{\theta})_{VR}$ raises concerns about the defensibility of this coefficient because systematic errors, in principle, should contribute to the true proficiency variance.

The final variance-ratio reliability coefficient, $R(\hat{\theta})_{MVR}$, is widely discussed and used in practice (e.g., Andersson & Xin, 2018; Cheng et al., 2012). Assuming the distribution of θ follows a standard normal distribution with a variance equal to 1, $R(\hat{\theta})_{MVR}$ can be expressed as $1/\{1 + \mathcal{E}[1/I(\theta, \hat{\theta})]\}$. The prevalent use of $R(\hat{\theta})_{MVR}$ relies on convenience—it does not require the use of sample estimates $\hat{\theta}$ s, assumes the variance of θ to be 1, and easily obtains the information function using standard normal quadrature points and weights. Coefficient $R(\hat{\theta})_{MVR}$ is an interesting statistic because it uses a mix of latent proficiency variance and expected error variance. Moreover, the sum of $\sigma^2(\theta)$ and $\mathcal{E}\sigma^2(\hat{\theta}|\theta)$ in the denominator is not equal to $\sigma^2(\hat{\theta})$, completely eliminating bias-related terms from the formula. Although it may be regarded as a more sensible coefficient than $R(\hat{\theta})_{VR}$, it has drawbacks because it engages the latent trait definition of proficiency, and the variance of actual proficiency estimates does not play any role.

It is interesting to note that $\sigma^2(\varepsilon)$ is never used as error variance for any of the variance-ratio reliability coefficients. According to the ANOVA identity, the overall error variance $\sigma^2(\varepsilon)$ can be formulated as (Lord, 1983):

$$\sigma^2(\varepsilon) = \sigma^2[\mathcal{E}(\varepsilon|\theta)] + \mathcal{E}\sigma^2(\varepsilon|\theta) = \sigma^2(BIAS) + \mathcal{E}\sigma^2(\hat{\theta}|\theta), \quad (37)$$

because $\mathcal{E}(\varepsilon|\theta) = \mathcal{E}(\hat{\theta} - \theta|\theta) = BIAS$ and $\sigma^2(\hat{\theta}|\theta) = \sigma^2(\varepsilon|\theta)$ for a single fixed test taker. Obviously, $\sigma^2(\varepsilon)$ includes both random error variance and variance attributable to bias, distinguishing it from the expected error variance $\mathcal{E}\sigma^2(\hat{\theta}|\theta)$ and the error variance defined in Equation 36.

Now, let us consider the parallel-forms coefficient, $R(\hat{\theta})_{PF} = \rho(\hat{\theta}, \hat{\theta}')$. This coefficient is, by definition, the Pearson correlation for a population between proficiency estimates obtained from two parallel forms. In cases where double testing is impractical, Lord (1983) derived a formula for $R(\hat{\theta})_{PF}$ that is expressed in terms of statistical quantities obtainable from a single administration of a test under the assumption of strictly parallel forms. Note that Samejima (1994) derived the same result as Lord (1983); however, she used the traditional assumptions of CTT, which may not hold exactly for $\hat{\theta}$. The parallel-forms reliability coefficient, $R(\hat{\theta})_{PF}$, is defined as

$$R(\hat{\theta})_{PF} \equiv \rho(\hat{\theta}, \hat{\theta}') = \frac{\sigma(\hat{\theta}, \hat{\theta}')}{\sigma(\hat{\theta})\sigma(\hat{\theta}')} = \frac{\sigma(\hat{\theta}, \hat{\theta}')}{\sigma^2(\hat{\theta})}, \quad (38)$$

where $\sigma(\hat{\theta}) = \sigma(\hat{\theta}')$ for strictly parallel forms. It can be shown that the asymptotically unbiased estimator of $\sigma(\hat{\theta}, \hat{\theta}')$ is

$$\sigma(\hat{\theta}, \hat{\theta}') = \sigma^2(\hat{\theta}) - \mathcal{E}\sigma^2(\hat{\theta}|\theta). \quad (39)$$

Therefore, the parallel-forms reliability coefficient of $\hat{\theta}$ becomes equal to $R(\hat{\theta})_{EVR}$:

$$\begin{aligned} R(\hat{\theta})_{PF} \equiv \rho(\hat{\theta}, \hat{\theta}') &= \frac{\sigma(\hat{\theta}, \hat{\theta}')}{\sigma^2(\hat{\theta})} = \frac{\sigma^2(\hat{\theta}) - \mathcal{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} \\ &= 1 - \frac{\mathcal{E}\sigma^2(\hat{\theta}|\theta)}{\sigma^2(\hat{\theta})} \equiv R(\hat{\theta})_{EVR}. \end{aligned} \quad (40)$$

The last coefficient to discuss is $R(\hat{\theta})_{SC}$, the squared-correlation reliability coefficient. S. Kim (2012) demonstrated that $R(\hat{\theta})_{SC}$ based on the linear regression of $\hat{\theta}$ on θ differs from $R(\hat{\theta})_{PF}$. Coefficient $R(\hat{\theta})_{SC}$ is defined as

$$R(\hat{\theta})_{SC} \equiv \rho^2(\hat{\theta}, \theta) = \frac{[\sigma(\hat{\theta}, \theta)]^2}{\sigma^2(\hat{\theta})\sigma^2(\theta)}, \quad (41)$$

where $\sigma(\hat{\theta}, \theta) = \sigma[\mathcal{E}(\hat{\theta}|\theta), \theta] + \mathcal{E}\sigma(\hat{\theta}, \theta|\theta) = \sigma^2(\theta) + \sigma(\theta, BIAS)$. Let $L(\varepsilon|\theta)$ denote the linear regression of ε on θ with the regression coefficient of $\beta(\varepsilon|\theta) = \sigma(\varepsilon, \theta)/\sigma^2(\theta) = \sigma(\theta, BIAS)/\sigma^2(\theta)$. It follows that

$$R(\hat{\theta})_{SC} = \frac{\sigma^2(\theta)}{\sigma^2(\hat{\theta})} [1 + \beta(\varepsilon|\theta)]^2 = R(\hat{\theta})_{VR} \times [1 + \beta(\varepsilon|\theta)]^2. \quad (42)$$

Because $\mathcal{E}(\varepsilon|\theta) = \mathcal{E}(\hat{\theta} - \theta|\theta) = BIAS \neq 0$, the slope of $L(\varepsilon|\theta)$, $\beta(\varepsilon|\theta)$, may not be zero. Then, it is clear from Equation 42 that $R(\hat{\theta})_{SC}$ and $R(\hat{\theta})_{VR}$ differ by a factor of $(1 + \beta_{\varepsilon|\theta})^2$. If $\sigma(BIAS, \theta)$ is positive, which usually is the case, $R(\hat{\theta})_{SC}$ will be larger than $R(\hat{\theta})_{VR}$. One way to estimate $R(\hat{\theta})_{SC}$ would be to use a mathematical expression for bias, which then can be used to estimate $\beta(\varepsilon|\theta)$. For example, Lord (1983) provided the bias function for $\hat{\theta}$ of θ under the three-parameter logistic model.

As evident from the formula itself, $R(\hat{\theta})_{SC}$ measures the extent to which proficiency estimates are associated with *latent* proficiencies. We can consider a similar (squared) correlation coefficient that involves *expected* proficiencies, denoted $\rho^2(\hat{\theta}, \mathcal{E}\hat{\theta})$. It is arguably more correct to refer to $\rho^2(\hat{\theta}, \mathcal{E}\hat{\theta})$ as a reliability coefficient, while $R(\hat{\theta})_{SC}$ can be seen as a squared validity coefficient. The key distinction lies in the fact that the two coefficients use definitions of true proficiency that differ by a constant *BIAS* for each test taker. If the bias is negligible, the distinction between the two will be trivial.

In many practical applications, $\hat{\theta}$ is often treated as an unbiased estimator, and the variance of the bias and its covariance with θ are neglected in estimating reliability. When the bias is negligibly small, such as for a long test, it may be argued that the expected variance terms are close to the variances associated with latent proficiencies: that is, $\sigma^2(\hat{\theta}) \cong \sigma^2(\theta) + \sigma^2(\varepsilon|\theta) \cong \sigma^2(\theta) + \mathcal{E}\sigma^2(\hat{\theta}|\theta) \cong \sigma^2[\mathcal{E}(\hat{\theta}|\theta)] + \mathcal{E}\sigma^2(\hat{\theta}|\theta)$.

Table 5.2 Variance-Ratio Reliability Coefficients for the Maximum Likelihood Estimator

Coefficient	Variance of Proficiency Estimates	True Proficiency Variance	Error Variance
$R(\hat{\theta})_{VR}$	$\sigma^2(\hat{\theta})$	$\sigma^2(\theta)$	$\mathcal{E}\sigma^2(\hat{\theta} \theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS)$
$R(\hat{\theta})_{EVR} = R(\hat{\theta})_{PF}$	$\sigma^2(\hat{\theta})$	$\sigma^2[\mathcal{E}(\hat{\theta} \theta)] = \sigma^2(\theta) + \sigma^2(BIAS) + 2\sigma(\theta, BIAS)$	$\mathcal{E}\sigma^2(\hat{\theta} \theta)$
$R(\hat{\theta})_{MVR}$	$\sigma^2(\theta) + \mathcal{E}\sigma^2(\hat{\theta} \theta)$	$\sigma^2(\theta)$	$\mathcal{E}\sigma^2(\hat{\theta} \theta)$

Consequently, the three variance-ratio reliability coefficients will yield similar results:

$R(\hat{\theta})_{VR} \cong R(\hat{\theta})_{EVR} \cong R(\hat{\theta})_{MVR}$. The parallel-forms reliability coefficient, $R(\hat{\theta})_{PF}$, has been proven to be equal to $R(\hat{\theta})_{EVR}$. Coefficient $R(\hat{\theta})_{SC}$ is limited to the linear regression assumption, and from the well-known statistical properties, $R(\hat{\theta})_{EVR}$ is generally larger than $R(\hat{\theta})_{SC}$. The two coefficients, $R(\hat{\theta})_{SC}$ and $R(\hat{\theta})_{EVR}$, will be identical only if bias is zero.

As highlighted by previous researchers (e.g., S. Kim, 2012; Lord, 1983), these coefficients are *not* estimating the same parameter and thus are not interchangeable. In general, the following inequalities are likely to hold: $R(\hat{\theta})_{PF} = R(\hat{\theta})_{EVR} \geq R(\hat{\theta})_{SC} \neq R(\hat{\theta})_{MVR} \geq R(\hat{\theta})_{VR}$. The fundamental cause of these differences is the bias in $\hat{\theta}$, and the extent to which these coefficients yield different results largely depends on the amount of bias. If it is suspected that bias is not negligible, the use of $R(\hat{\theta})_{EVR}$ would be advisable. Table 5.2 summarizes the variance terms involved in each of the variance-ratio reliability coefficients.

The impact of bias in reliability coefficients can be alleviated by using modified ML estimators that adjust for the bias of $\hat{\theta}$. For example, Lord (1983) derived a bias-correction function for $\hat{\theta}$ under the three-parameter logistic model. Samejima (1993a, 1993b) expanded Lord's (1983) work to develop a bias function for any discrete item responses. Warm's (1989) weighted likelihood estimator is also known to reduce the bias of $\hat{\theta}$. However, in general, there is a dearth of literature focused on the effect of bias in reliability for various IRT models, including polytomous and multidimensional models. So, caution needs to be exercised when deciding which formula to use in practice.

Bayesian Proficiency Estimates

The Bayesian proficiency estimator of primary interest in this section is the EAP estimator denoted $\tilde{\theta}$. A crucial difference between the ML and EAP estimators in conceptualizing reliability lies in the use of different conditional distributions. The ML estimator focuses on the distribution of $f(\hat{\theta}|\theta)$, while the EAP estimator is concerned with $f(\theta|\tilde{\theta})$. The variance of $f(\theta|\tilde{\theta})$ serves as the conditional error variance for EAP estimates, denoted $\sigma^2(\theta|\tilde{\theta})$. The commonly discussed and utilized reliability coefficients for $\tilde{\theta}$ include:

$$\begin{aligned}
R(\tilde{\theta})_{SC} &\equiv \rho^2(\theta, \tilde{\theta}); \\
R(\tilde{\theta})_{VR} &\equiv \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\theta)} = \frac{\sigma^2[\mathcal{E}(\theta|\tilde{\theta})]}{\sigma^2(\theta)} = 1 - \frac{\mathcal{E}\sigma^2(\theta|\tilde{\theta})}{\sigma^2(\theta)} = \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\tilde{\theta}) + \mathcal{E}\sigma^2(\theta|\tilde{\theta})}; \text{ and} \\
R(\tilde{\theta})_{INF} &\equiv 1 - \frac{\mathcal{E}[1/(I(\theta, \tilde{\theta}) + 1)]}{\sigma^2(\theta)}.
\end{aligned}$$

The parallel-forms definition of reliability is *not* applicable to $\tilde{\theta}$ in the Bayesian framework because the obtained data are considered fixed. The squared-correlation definition of $R(\tilde{\theta})_{SC}$ resembles that of $R(\hat{\theta})_{SC}$ for the ML estimator, differing in that $R(\tilde{\theta})_{SC}$ is based on the linear regression of θ and $\tilde{\theta}$, $L(\theta|\tilde{\theta})$, while $R(\hat{\theta})_{SC}$ depends on $L(\hat{\theta}|\theta)$. Unlike the ML estimator, all variations of the variance-ratio coefficients associated with $R(\tilde{\theta})_{VR}$ are identical. Notably, the reversal of variances in true proficiencies and proficiency estimates in $R(\tilde{\theta})_{VR}$ compared to the ML estimator stems from the Bayesian framework's focus on predicting θ given $\tilde{\theta}$, as opposed to estimating $\hat{\theta}$ given θ under the ML framework. The last coefficient, $R(\tilde{\theta})_{INF}$, is proposed as an alternative reliability coefficient for $\tilde{\theta}$, approximating the relationship between the test information function (noted by the *INF* subscript) and the variance of the posterior distribution. As an approximation to $R(\tilde{\theta})_{VR}$, $R(\tilde{\theta})_{INF}$ offers the benefit of avoiding intricate computations of posterior variance while integrating the well-established concept of the test information function.

The EAP estimate for a test taker with item response data $\mathbf{u} = \{u_1, \dots, u_n\}$, by definition, is the mean of the posterior distribution, which is given by

$$\tilde{\theta} = \mathcal{E}(\theta|\mathbf{u}) = \frac{\int_{\theta} \theta \prod_{i=1}^n \Pr(U_i = u_i|\theta) f(\theta) d\theta}{\int_{\theta} \prod_{i=1}^n \Pr(U_i = u_i|\theta) f(\theta) d\theta}, \quad (43)$$

where $\Pr(U_i = u_i|\theta)$ is the item response function for a given IRT model and $f(\theta)$ is the prior distribution of θ , which often is assumed to be the same for all test takers. The variance of the posterior distribution (i.e., error variance) for a test taker with $\mathbf{u}$ is (Bock & Mislevy, 1982):

$$\sigma^2(\theta|\mathbf{u}) = \frac{\int_{\theta} (\theta - \tilde{\theta})^2 \prod_{i=1}^n \Pr(U_i = u_i|\theta) f(\theta) d\theta}{\int_{\theta} \prod_{i=1}^n \Pr(U_i = u_i|\theta) f(\theta) d\theta}. \quad (44)$$

The integrals in Equations 43 and 44 are replaced by summations using a discrete proficiency distribution for computational purposes.

Of note, one distinct property of $\tilde{\theta}$ is its one-to-one correspondence with $\mathbf{u}$, meaning that test takers with the same response pattern will have the same $\tilde{\theta}$. This property suggests that the conditional posterior distribution of θ is identical using either $\mathbf{u}$ or $\tilde{\theta}$

as the conditioning variable. As such, $\tilde{\theta} = \mathcal{E}(\theta|\mathbf{u}) = \mathcal{E}(\theta|\tilde{\theta})$ and $\sigma^2(\theta|\mathbf{u}) = \sigma^2(\theta|\tilde{\theta})$. In the context of predicting θ given $\tilde{\theta}$, the ANOVA identity suggests that

$$\sigma^2(\theta) = \sigma^2[\mathcal{E}(\theta | \tilde{\theta})] + \mathcal{E}\sigma^2(\theta | \tilde{\theta}). \quad (45)$$

Based on this equality, the intraclass correlation coefficient is $\sigma^2[\mathcal{E}(\theta|\tilde{\theta})]/\sigma^2(\theta) = 1 - [\mathcal{E}\sigma^2(\theta|\tilde{\theta})/\sigma^2(\theta)]$.

One appealing feature of the Bayesian framework is that the distinction of expected variance terms is no longer necessary. The first term in the right-hand side of Equation 45, $\sigma^2[\mathcal{E}(\theta | \tilde{\theta})]$, is equal to $\sigma^2(\tilde{\theta})$ because $\mathcal{E}(\theta | \tilde{\theta}) = \tilde{\theta}$. The equality of $\sigma^2[\mathcal{E}(\theta | \tilde{\theta})] = \sigma^2(\tilde{\theta})$ is sufficient to render the first three variance-ratio definitions in $R(\tilde{\theta})_{VR}$ identical. It can be further shown that $\sigma^2(\varepsilon) = \sigma^2[\mathcal{E}(\varepsilon|\tilde{\theta})] + \mathcal{E}\sigma^2(\varepsilon|\tilde{\theta})$, similar to Equation 37, and because $\sigma^2[\mathcal{E}(\varepsilon|\tilde{\theta})] = 0$, $\sigma^2(\varepsilon) = \mathcal{E}\sigma^2(\varepsilon|\tilde{\theta}) = \mathcal{E}\sigma^2(\theta|\tilde{\theta})$. As a result,

$$\sigma^2(\theta) = \sigma^2(\tilde{\theta}) + \sigma^2(\varepsilon) = \sigma^2(\tilde{\theta}) + \mathcal{E}\sigma^2(\varepsilon | \tilde{\theta}). \quad (46)$$

It follows that various variance-ratio definitions ascribed to $R(\tilde{\theta})_{VR}$ are all equal to one another.

Similar to Equation 42 for the ML estimator, the EAP version of the squared-correlation coefficient is given by

$$R(\tilde{\theta})_{SC} \equiv \frac{\sigma^2[L(\theta|\tilde{\theta})]}{\sigma^2(\theta)} = \frac{\sigma^2(\tilde{\theta})}{\sigma^2(\theta)}. \quad (47)$$

Equation 47 clearly suggests that $R(\tilde{\theta})_{SC} = R(\tilde{\theta})_{VR}$. So, for the EAP estimator, $R(\tilde{\theta})_{SC}$ and various alternative definitions associated with $R(\tilde{\theta})_{VR}$ are all equal. This equality occurs as a result of the following one-to-one correspondence: $\mathcal{E}(\theta|\mathbf{u}) = \mathcal{E}(\theta|\tilde{\theta}) = \tilde{\theta}$. If the one-to-one correspondence is absent for an estimator, the equality may not hold. For example, S. Kim (2012) cautioned that another Bayes estimator, MAP, does not necessarily exhibit the one-to-one correspondence with the one- or two-parameter logistic IRT models.

The coefficient $R(\tilde{\theta})_{INF}$ incorporates the concept of information, typically associated with the ML estimator, to estimate the conditional and overall error variances. The link between the information for $\tilde{\theta}$ and $\sigma^2(\theta|\tilde{\theta})$ is established by the argument that the information provided by the prior distribution of θ is equivalent to adding an item to which all test takers in the population respond identically (Ferrando & Lorenzo-Seva, 2007; Thissen & Orlando, 2001). If the prior distribution of θ follows the standard normal distribution, as is commonly assumed, its contribution to the test information function is a constant value equal to 1 because $1/\sigma^2(\theta) = 1$. Consequently,

$$\sigma_{INF}^2(\theta | \tilde{\theta}) \cong \frac{1}{I(\theta, \tilde{\theta}) + 1}, \quad (48)$$

where $I(\theta, \tilde{\theta})$ is the traditional test information for a given θ . Furthermore, $\mathcal{E}\sigma^2(\theta|\tilde{\theta})$ in $R(\tilde{\theta})_{VR}$ can be replaced with $\sigma_{INF}^2(\theta|\tilde{\theta})$ computed by Equation 48 to derive an alternative expression for the reliability coefficient for $\tilde{\theta}$:

$$R(\tilde{\theta})_{INF} = 1 - \frac{\mathcal{E}[1/(I(\theta, \hat{\theta}) + 1)]}{\sigma^2(\theta)}, \quad (49)$$

which can be further reduced to $R(\tilde{\theta})_{INF} = 1 - \mathcal{E}[1/(I(\theta, \hat{\theta}) + 1)]$ because $\sigma^2(\theta)$ is equal to 1.

It is fair to say that coefficient $R(\tilde{\theta})_{INF}$ is computationally more convenient and capitalizes on the useful concept of information concerning its error variance. However, $R(\tilde{\theta})_{INF}$ would function merely as an approximation to the more defensible coefficient $R(\tilde{\theta})_{VR}$. This is because the algebraic similarity in aggregate statistics does not necessarily imply that the interpretation of error variance and its relation to other variance terms, especially in terms of conditional statistics, remains unchanged.

Summed Raw Scores and Transformed Scale Scores

IRT can serve as an underlying psychometric framework for estimating reliability for both summed raw scores, denoted X , and transformed scale scores, denoted $S = t(X)$. The general procedures suggested in the literature on this topic are largely consistent, except for variations in presentation details (e.g., S. Kim & Feldt, 2010; Kolen et al., 1996; Lord, 1980). As a function of a proficiency distribution and item parameters, the probability distributions for all three components of a measurement model (i.e., observed, true, and error scores) can be modeled and used subsequently to form reliability statistics.

A general model for representing marginal observed scores is given by

$$f(x) = \Pr(X = x) = \int_{\theta} \Pr(X = x|\theta) f(\theta) d\theta, \quad (50)$$

where $\Pr(X = x | \theta)$ is the conditional raw-score distribution and $f(\theta)$ is the distribution of θ . The conditional distribution, $f(x|\theta) = \Pr(X = x|\theta)$, captures the score variability for a test taker with θ over repeated testing using the strictly parallel forms or a fixed form. It can be efficiently computed using a recursive formula (e.g., Hanson, 1994; Lord & Wingersky, 1984; Thissen et al., 1995). Assuming a specific θ distribution (e.g., standard normal), the marginal observed score distribution, $f(x)$, is obtained by integrating the conditional distributions. The resulting $f(x)$ is often called the fitted or model-based observed score distribution, which lends itself to many applications in measurement such as test score equating, model fit examination, score smoothing, and quantifying reliability statistics.

For a test taker with θ , the expected value of $f(x|\theta)$ is the IRT analogue of true score in CTT, denoted

$$\tau_{\theta} = \mathcal{E}(X|\theta) = \sum x \Pr(X = x|\theta), \quad (51)$$

where the summation is taken over all possible values of X . The IRT true score τ_{θ} , a monotone nonlinear transformation of θ , is also referred to as the TCC. The variance of $\Pr(X = x|\theta)$ is the conditional error variance, computed by

$$\sigma^2(X|\theta) = \sum x^2 \Pr(X = x|\theta) - \tau_{\theta}^2. \quad (52)$$

The expected value of $\sigma^2(X|\theta)$ over all test takers in the population is the overall error variance, that is, $\mathcal{E}\sigma^2(X|\theta) = \int_{\theta} \sigma^2(X|\theta) f(\theta) d\theta$.

The ANOVA decomposition of the observed score variance is given by

$$\sigma^2(X) = \sigma^2[\mathcal{E}(X|\theta)] + \mathcal{E}\sigma^2(X|\theta) = \sigma^2(\tau_{\theta}) + \mathcal{E}\sigma^2(X|\theta). \quad (53)$$

Based on this ANOVA identity, the most widely used reliability coefficient is the following intraclass correlation:

$$R(X)_{IRT} = \frac{\sigma^2(\tau_{\theta})}{\sigma^2(\tau_{\theta}) + \mathcal{E}\sigma^2(X|\theta)} = 1 - \frac{\mathcal{E}\sigma^2(X|\theta)}{\sigma^2(X)}. \quad (54)$$

S. Kim and Feldt (2010) referred to the first formula as the true score variance approach and the second formula as the observed score variance approach. Andersson and Xin (2018) called $R(X)_{IRT}$ the test reliability. The true score variance, $\sigma^2(\tau_{\theta})$, in Equation 54 can be obtained using

$$\sigma^2(\tau_{\theta}) = \int_{\theta} \tau_{\theta}^2 f(\theta) d\theta - (\mathcal{E}\tau_{\theta})^2, \quad (55)$$

where $\mathcal{E}\tau_{\theta} = \int_{\theta} \tau_{\theta} f(\theta) d\theta$. The second formula of $R(X)_{IRT}$ requires observed score variance, which is the variance of $f(x)$ in Equation 50. If the sample observed score variance, denoted $\hat{\sigma}^2(X)$, is used, the variance equality in Equation 53 is lost, and the two formulas of $R(X)_{IRT}$ are no longer exchangeable. The coefficient, $1 - [\mathcal{E}\sigma^2(X|\theta)/\hat{\sigma}^2(X)]$, might be best viewed as a sample estimate of $R(X)_{IRT}$, although there is no clear theoretical justification for doing so. The sample estimate may work well if the model fit is reasonably good and may sometimes be the only option when the model-based observed score variance cannot be computed easily (e.g., noninteger observed scores).

Note that there is a one-to-one correspondence between θ and τ_{θ} —that is, $\mathcal{E}(X|\theta) = \mathcal{E}[X|\tau_{\theta}] = \tau_{\theta}$. As a result, the squared-correlation coefficient $\rho^2(X, \tau)$ will be identical to $R(X)_{IRT}$.

Now let us consider scale scores, $S = t(X)$, that are transformed from raw scores. According to the well-known statistical property of correlation, a linear transformation of variables does not alter the value of correlation, assuming the direction of transformation for both variables is the same. This indicates that any linearly transformed scale scores will have the same reliability as that for raw scores. Thus, linear transformation is not discussed further in this section. By contrast, nonlinear transformation requires special treatment, as discussed next.

Suppose a conversion table contains all possible raw-score points, each corresponding to a scale-score point with a nonlinear relationship. The raw-to-scale score conversion table can be one-to-one or many-to-one and may contain integer or noninteger scale scores. Conceptually, over repeated measurements, the probability of a test taker receiving a certain scale score is the same as the probability of obtaining the corresponding

raw score—that is, $\Pr[t(x)|\theta] = \Pr(X = x|\theta)$. The expected value of scale scores over replications for a test taker with θ is called *true scale score* for the test taker and is given by

$$\xi_\theta = \mathcal{E}(S|\theta) = \sum t(x)\Pr(X = x|\theta), \quad (56)$$

where the summation is over the entire range of raw scores. Likewise, the variance of scale scores for given θ is defined as

$$\sigma^2(S|\theta) = \sum [t(x)]^2 \Pr(X = x|\theta) - \xi_\theta^2. \quad (57)$$

The square root of Equation 57 is the *conditional scale-score SEM*. A reliability coefficient for scale scores can be expressed as

$$R(S)_{IRT} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(\xi_\theta) + \mathcal{E}\sigma^2(S|\theta)} = 1 - \frac{\mathcal{E}\sigma^2(S|\theta)}{\sigma^2(S)}, \quad (58)$$

where $\mathcal{E}\sigma^2(S|\theta) = \int \sigma^2(S|\theta)g(\theta)d\theta$ and $\sigma^2(\xi_\theta)$ can be obtained similarly to Equation 55.

Composite Scores and Composite Scale Scores

A composite score, as defined earlier, is a weighted sum of scores from multiple constituent components. The potential score points and range of the composite score depend on the combinations of component scores and weights. In practical applications, composite scores are often rounded to the nearest integer, especially when noninteger weights are employed. However, rounding is not obligatory as long as all possible composite scores can be accurately identified and modeled.

The framework presented in this section closely aligns with the work of W. Lee et al. (2020), and similar developments can be found in Kolen et al. (2012) and Kolen and Lee (2011). W. Lee et al. (2020) considered three underlying IRT frameworks: unidimensional IRT (UIRT), simple-structure MIRT (SS-MIRT), and bifactor MIRT (BF-MIRT) models, each addressing potential multidimensionality in different ways. While the assumption of unidimensionality may be somewhat violated for composite scores, UIRT models are still commonly used. In the SS-MIRT model, each item is assumed to measure a single construct (i.e., component) and constructs across multiple components can be correlated (Kolen et al., 2012; W. Lee et al., 2020). The BF-MIRT model (Gibbons & Hedeker, 1992; Gibbons et al., 2007) assumes that all items in the test measure the same general construct, and each item measures an additional construct specific to the associated component. The general and specific constructs are typically uncorrelated. Further information about BF-MIRT can be found in Cai et al. (2011) and DeMars (2006, 2013).

The modeling of marginal composite scores follows a structure similar to Equation 50, with the use of a vector θ as the conditioning variable:

$$f(z) = \Pr(Z = z) = \int_{\theta} \Pr(Z = z|\theta)f(\theta)d\theta. \quad (59)$$

The number of proficiency parameters in the vector θ depends on the model—for UIRT models, $\theta = \theta$; for the SS-MIRT model, $\theta = \{\theta_1, \dots, \theta_M\}$; and for the BF-MIRT model,

$\theta = \{\theta_G, \theta_1, \dots, \theta_M\}$, where θ_G indicates the general factor. Numerical integration can be carried out using a univariate standard normal distribution for UIRT or a multivariate normal density function for the MIRT models, with nonzero correlations between components for the SS-MIRT or zero correlations (typically) for the BF-MIRT.

Assuming conditional local independence, one way to compute the conditional composite score distribution, $f(z|\theta) = \Pr(Z = z|\theta)$, is

$$\Pr(Z = z|\theta) = \sum_{Z=w_1X_1+\dots+w_MX_M} \prod_{m=1}^M \Pr(X_m = x_m|\theta), \quad (60)$$

where the summation is taken over all possible combinations of weighted component scores that lead to the same composite score. The conditional distribution for each component can be computed using a recursion formula such as Lord and Wingersky's (1984) method. Note that not all elements of θ are used in computing the conditional distribution for each component. For example, for component m , only a single θ_m is involved for the SS-MIRT model, while both θ_G and θ_m are needed for the BF-MIRT model. The mean (i.e., true composite score) and variance of the conditional distribution, respectively, are

$$\zeta_\theta = \mathcal{E}(Z|\theta) = \sum z \Pr(Z = z|\theta), \quad (61)$$

and

$$\sigma^2(Z|\theta) = \sum z^2 \Pr(Z = z|\theta) - \zeta_\theta^2. \quad (62)$$

A reliability coefficient then is given by

$$R(Z)_{IRT} = \frac{\sigma^2(\zeta_\theta)}{\sigma^2(\zeta_\theta) + \mathcal{E}\sigma^2(Z|\theta)} = 1 - \frac{\mathcal{E}\sigma^2(Z|\theta)}{\sigma^2(Z)}. \quad (63)$$

Alternatively, $R(Z)_{IRT}$ can be expressed in the form of the general reliability-coefficient formula for linear composite scores (i.e., Equation 15), which involves error variances for individual components. That is,

$$R(Z)_{IRT^*} = 1 - \frac{\sum_{m=1}^M \mathcal{E}\sigma^2(X_m|\theta)}{\sigma^2(Z)}, \quad (64)$$

where $\mathcal{E}\sigma^2(X_m|\theta)$ is the overall error variance for each component and errors across components are assumed to be uncorrelated. Comparing $R(Z)_{IRT}$ and $R(Z)_{IRT^*}$, it is evident that computing each $\mathcal{E}\sigma^2(X_m|\theta)$ and adding them up is easier than computing $\mathcal{E}\sigma^2(Z|\theta)$, which relies on the computationally cumbersome conditional composite score distribution, $f(z|\theta)$. However, $\sigma^2(Z)$ in the denominator of both coefficients is based on $f(z)$ in Equation 59, which, in turn, requires the use of $f(z|\theta)$. Therefore, bypassing $f(z|\theta)$ in the computational process is not a viable option. If the sample composite score variance, $\hat{\sigma}^2(Z)$, is used in place of the model-based one

for $R(Z)_{IRT^*}$, some of the computational complexity will be alleviated. Doing so, however, will result in a reliability estimate that is best viewed as a sample estimate of $R(Z)_{IRT^*}$ (or $R(Z)_{IRT}$) and many desired features of the coherent psychometric framework will no longer be available. Yet, the computational simplicity might be very useful in practice.

Composite scores usually are further transformed nonlinearly to scale scores for reporting purposes. The results for composite scale scores, $S = t(Z)$, can be obtained in an analogous manner by noting that $\Pr[t(z)|\theta] = \Pr(Z = z|\theta)$ and $\Pr[t(z)] = \Pr(Z = z)$. Replacing Z with $t(z)$ in Equations 61, 62, and 63 yields the true score, conditional error variance, and reliability coefficient all on the composite scale score metric.

Scale Scores Transformed From Proficiency Estimates

In situations where IRT proficiency scoring is used, the proficiency estimates derived from ML or EAP estimation methods are commonly converted to scale scores for reporting purposes. Here, we are mainly concerned with nonlinear transformations. Let $t(\theta)$ be a monotone increasing transformation function that converts θ to scale scores. It is assumed that the transformation is applied to both true and estimated proficiencies such that $S = t(\hat{\theta})$ and $\mathcal{E}(S|\theta) = \xi_\theta = t(\theta)$. The transformation function is further assumed to be continuous and differentiable at every level of θ . In cases where a mathematical expression for a theta-to-scale score transformation is not available, a practical solution might involve employing a smoothing technique to obtain a smooth continuous conversion relationship, such as a high-degree polynomial.

Let us begin with the ML estimator. According to Lord (1980, p. 67), the test information of observed scale score, S , for making inferences about θ , is defined as

$$I(\theta, S) = \frac{\left[\frac{\partial}{\partial \theta} \mathcal{E}(S|\theta) \right]^2}{\sigma^2(S|\theta)} = \frac{\left[t'(\theta) \right]^2}{\sigma^2(S|\theta)}, \quad (65)$$

where $t'(\theta)$ is the first derivative function of $t(\theta)$ and

$$\sigma^2(S|\theta) \cong \left[t'(\theta) \right]^2 \sigma^2(\hat{\theta}|\theta). \quad (66)$$

Equation 66 assumes that $\hat{\theta}$ is an asymptotically unbiased estimator of θ and is known as the delta method approximation (Kendall & Stuart, 1977), which will be detailed in the next section. Equations 65 and 66 suggest that $I(\theta, S) = I(\theta, \hat{\theta})$. Furthermore, the monotone increasing relationship between θ and $\xi_\theta \equiv \mathcal{E}(S|\theta)$ implies that $\sigma^2(S|\theta) = \sigma^2(S|\xi_\theta)$. The overall scale-score error variance can be obtained by integrating $\sigma^2(S|\theta)$ over the proficiency distribution as

$$\mathcal{E}\sigma^2(S|\theta) \cong \mathcal{E} \left(\frac{\left[t'(\theta) \right]^2}{I(\theta, S)} \right) = \mathcal{E} \left(\frac{\left[t'(\theta) \right]^2}{I(\theta, \hat{\theta})} \right). \quad (67)$$

The variance of true scale scores is

$$\sigma^2(\xi_\theta) = \int_{\theta} \xi_\theta^2 f(\theta) d\theta - (\mathcal{E}\xi_\theta)^2, \quad (68)$$

which can be evaluated using numerical integration. The scale-score analogues of $R(\hat{\theta})_{VR}$, $R(\hat{\theta})_{EVR}$, and $R(\hat{\theta})_{MVR}$, respectively, are

$$R(S_{\hat{\theta}})_{VR} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(S)}; \quad R(S_{\hat{\theta}})_{EVR} = 1 - \frac{\mathcal{E}\sigma^2(S|\theta)}{\sigma^2(S)}; \text{ and}$$

$$R(S_{\hat{\theta}})_{MVR} = \frac{\sigma^2(\xi_\theta)}{\sigma^2(\xi_\theta) + \mathcal{E}\sigma^2(S|\theta)}.$$

Similar to ML estimates, these three coefficients for scale scores do not necessarily yield the same result.

For the EAP estimator, the conditional error variance, $\sigma^2(\theta|\mathbf{u}) = \sigma^2(\theta|\tilde{\theta})$, is provided in Equation 44. The conditional error variance for scale scores can be found using the delta method approximation as

$$\sigma^2(\xi_\theta | \tilde{\theta}) \cong \left\{ t' [\mathcal{E}(\theta|\tilde{\theta})] \right\}^2 \sigma^2(\theta|\tilde{\theta}) = \left[t'(\tilde{\theta}) \right]^2 \sigma^2(\theta|\tilde{\theta}). \quad (69)$$

Note that, under the Bayesian framework, $\sigma^2(\xi_\theta | \tilde{\theta})$ is the variance of *true* scale scores conditional on $\tilde{\theta}$. It immediately follows that the overall scale-score error variance is $\mathcal{E}\sigma^2(\xi_\theta|\tilde{\theta}) \cong \mathcal{E} \left\{ \left[t'(\tilde{\theta}) \right]^2 \sigma^2(\theta|\tilde{\theta}) \right\}$, where the expectation (i.e., average) is taken over all persons in the data. A reliability coefficient for scale scores takes a form similar to $R(\tilde{\theta})_{VR}$:

$$R(S_{\hat{\theta}})_{VR} = \frac{\sigma^2(S)}{\sigma^2(\xi_\theta)} \cong 1 - \frac{\mathcal{E}\sigma^2(\xi_\theta|\tilde{\theta})}{\sigma^2(\xi_\theta)} \cong \frac{\sigma^2(S)}{\sigma^2(S) + \mathcal{E}\sigma^2(\xi|\tilde{\theta})}, \quad (70)$$

where the true scale-score variance, $\sigma^2(\xi_\theta)$, can be computed using Equation 68. Unlike the results for $\tilde{\theta}$, the three variations of $R(S_{\hat{\theta}})_{VR}$ may not yield the same results because of the use of the delta method approximation and possible rounding of reported scale scores.

As discussed earlier, the Bayesian conditional error variance can also be approximated using the test information function as $\sigma_{INF}^2(\theta | \tilde{\theta}) \cong 1 / [I(\theta, \theta) + 1]$. In this approximation, the conditioning variable becomes true θ and a discrete set of quadrature points can be used instead of individual $\tilde{\theta}$'s. Then, the conditional scale-score variance can be estimated using the delta method, which gives $\sigma^2(\xi_\theta|\theta) \cong \left[t'(\theta) \right]^2 \sigma_{INF}^2(\theta|\tilde{\theta})$, and the overall error variance is obtained by integrating $\sigma^2(\xi_\theta|\theta)$ over the θ quadrature distribution. Finally, the scale-score version of $R(\tilde{\theta})_{INF}$ is given by $R(S_{\hat{\theta}})_{INF} = 1 - \left[\mathcal{E}\sigma^2(\xi_\theta|\theta) / \sigma^2(\xi_\theta) \right]$.

ESTIMATORS OF CSEMs

Conceptually, SEM represents the standard deviation of test scores over repeated measurements for a person, assuming the measurement process does not alter their true score. In CTT, the SEM is commonly expressed as $\sigma(E) = \sigma(X)\sqrt{1-\rho(X, X')}$, referred to as the overall SEM. While this computation offers simplicity, an excessive focus on reliability coefficients may divert attention from actively considering whether the sources of error involved in the SEM closely align with the intended use and generalization of the test results. Additionally, applying the same SEM to all test takers when constructing confidence intervals overlooks variations in SEM magnitude as a function of true and observed scores. An arguably more defensible approach might involve starting with CSEMs, which can then be used to derive both the overall SEM [see Equation 6] and reliability coefficients, if needed.

Although direct estimation of CSEMs over many repeated measurements is practically unfeasible, collecting data from two replications may not be so unrealistic. This approach is similar to the data collection design used for estimating the test-retest and parallel-forms reliability coefficients. Using two sets of observed scores, X_1 and X_2 , an estimate of the CSEM for a single test taker can be computed directly as

$$CSEM_R = \frac{|X_1 - X_2|}{\sqrt{2}}. \quad (71)$$

Some benefits of Equation 71 were argued by Brennan (2001a), including that, by replicating the full-length measurement procedure twice, the investigator is forced to focus on the actual characteristics of replications to ensure that they are faithful reflections of the intended universe. In addition, Equation 71 provides a direct estimate of the CSEM without making any assumptions, applicable to various score types such as raw scores, scale scores, and composite scores. This is especially advantageous given that computing CSEMs for scale scores based on a single administration of a test can often be complicated (e.g., Kolen et al., 1992, 1996; W. Lee et al., 2000). Despite these potential benefits, Equation 71 is not widely used in practice, partly because of the questionable accuracy of estimation based on only two replications and the challenge of obtaining two full-length replications on adequate samples.

Similar to reliability coefficients, considerable effort has been invested in developing estimators for CSEMs based on a single test administration using various psychometric models and making assumptions that are often too strong or untestable. In theory, CSEM is the standard deviation of observed scores over repeated parallel (classically, randomly, etc.) measurements conditional on each person's true score: that is, $\sigma(X|\tau_p) = \sigma(E|\tau_p)$. (Note that the subscript p denotes a particular person but may be omitted for simplicity when context allows.) Estimators for CSEMs fall broadly into two categories: those that condition on observed scores and those that condition on true scores. The former category avoids assumptions about the true score distribution, employing each test taker's observed score as an estimator of their true score, $\hat{\tau}_p = x_p$, since true scores are unknown.

Lord (1984) and Woodruff (1990) noted, however, that considering CSEM as conditional on observed score is appropriate only when a test taker at a particular observed score is randomly selected. Nonetheless, the bias in estimated CSEMs using observed scores as a conditioning variable diminishes with increasing test reliability. Also, studies have found this bias to be practically minimal (Feldt et al., 1985; W. Lee et al., 2000; Woodruff et al., 2013).

The second type of procedures make certain assumptions to estimate the true score distribution and express CSEMs conditional on true score. Consequently, these procedures eliminate the bias issue and ambiguity of interpretation. However, if reporting the CSEM to each individual test taker is the goal, mapping observed scores to true scores used for CSEM computation becomes inevitable, which makes the practice essentially the same as using observed scores as a conditioning variable. Both types of procedures generally yield similar results, and no particular argument is made here for preferring one approach over the other. Rather, it is advised that the choice of method should depend on the characteristics of available data, reasonableness of assumptions underlying the psychometric model, and available computer resources.

Summed Raw Scores

Thorndike Method

Thorndike (1951) was one of the first investigators to examine CSEMs that vary as a function of observed scores based on split-half tests. Suppose the total test score X is split into two essentially tau-equivalent half test scores, $X_1 = (1/2)T + C_1 + E_1$ and $X_2 = (1/2)T + C_2 + E_2$, and $X = X_1 + X_2$. It can be shown that, under CTT assumptions, the variance of difference scores between half-test scores is equal to the error variance for the total test:

$$\sigma^2(X_1 - X_2) = \sigma^2(C_1 - C_2 + E_1 - E_2) = \sigma^2(E_1) + \sigma^2(E_2) = \sigma^2(E). \quad (72)$$

This implies that, for any group of test takers with total observed score x , the standard deviation of difference scores between X_1 and X_2 provides an estimate of the CSEM for that group:

$$CSEM(X)_{TH} = \sigma(X_1 - X_2 | x). \quad (73)$$

However, a limitation arises as CSEMs across the range of total scores can be erratic, particularly at extreme scores due to low test-taker frequencies. In practice, test takers can be grouped into short intervals based on their total scores, and the CSEM is computed for each interval (i.e., subgroup).

Mollenkopf Method

Mollenkopf (1949) developed an estimator for CSEMs based on split halves, often considered a refinement of Thorndike's method (Feldt & Qualls, 1996). Assuming tau equivalence of the two split-half tests, consider the following quantity for person p :

$$D_p = [(X_{p1} - X_{p2}) - (\bar{X}_1 - \bar{X}_2)]^2, \quad (74)$$

where $\bar{X}_1$ and $\bar{X}_2$ are group means on the two half-tests. For any subgroup of N test takers, the average of D_p values, by definition, is the variance of half-test difference scores, which is equal to the error variance for the full-length test:

$$\bar{D} = \frac{1}{N} \sum_{p=1}^N [(X_{p1} - X_{p2}) - (\bar{X}_1 - \bar{X}_2)]^2 = \sigma^2(X_1 - X_2) = \sigma^2(E). \quad (75)$$

The square root of $\bar{D}$ for a subgroup of test takers with an observed score x is the CSEM, which is identical to Thorndike's CSEM. Mollenkopf (1949) goes further to suggest using a polynomial regression with D_p in Equation 74 as a dependent variable and the total score X as an independent variable:

$$\hat{D} = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_q X^q, \quad (76)$$

where the polynomial degree q is chosen by the investigator. If the regression model fits well, the resulting predicted values, $\hat{D}$, provide a better approximation of conditional means than $\bar{D}$. Once the polynomial regression coefficients are estimated with a chosen q , Mollenkopf's CSEM at observed score x is

$$CSEM(X)_{MO} = \sqrt{\hat{D}} = \sqrt{\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q}. \quad (77)$$

Although fitting a polynomial regression is cumbersome, $CSEM(X)_{MO}$ has the desirable characteristic of providing smoother patterns of CSEMs along the observed score scale.

Binomial Error Model

A notable drawback of $CSEM(X)_{TH}$ and $CSEM(X)_{MO}$ is that they require splitting the full test into two parallel halves, and there are various possibilities for obtaining such splits. Lord's (1955, 1957) seminal work on the binomial error model provided a substantially different perspective. In this framework, a test form with n dichotomously scored items is viewed as a random sample from a universe of such items (i.e., each item has an equal probability of being selected). Each test taker has a true proportion-correct score, $\pi_p = \tau_p/n$, such that the probability of answering each undifferentiable item correctly is π_p for test taker p . Measurement errors are conceptualized by the variation in observed scores across an infinite number of test forms, each containing a different set of randomly selected n items.

For a test taker with π_p , the distribution of observed scores from repeated measurements follows a binomial distribution with the parameter π_p , and the error variance for the test taker is $\sigma^2(E|\pi_p) = \sigma^2(X|\pi_p) = n\pi_p(1-\pi_p)$. Lord's CSEM, replacing the unknown parameter π_p with the observed proportion-correct score $\bar{x}_p = x_p/n$, is given by

$$CSEM(X)_{BN} = \hat{\sigma}(E|\bar{x}_p) = \sqrt{\frac{x_p(n-x_p)}{n-1}}. \quad (78)$$

This formula can compute CSEMs for raw-score points ranging from 0 to n without needing test-taker data. It has a quadratic form, ensuring the resulting CSEMs exhibit a smooth inverted-U shape with zero values at observed scores of zero and n .

Criticism of Lord's CSEM involves its failure to differentiate items, treating them as having equal difficulties under the binomial error model. In response, Keats (1957) proposed a correction, introducing a fuzzy factor of $(1 - KR20)/(1 - KR21)$ to multiply the squared values of Lord's CSEMs. This adjustment is intended to bring the overall error variance, when averaged, into alignment with the error variance in KR20. From the perspective of CTT, the criticism and Keats's adjustment find support. However, from the viewpoints of GT, Lord's model possesses valuable features that were not clearly identifiable at the time. The notion of random sampling of items over replications diverges from CTT definitions of parallelism but aligns with the key characteristic of GT—that is, the randomly parallel forms assumption. Moreover, Lord's error variance can be shown to be identical to the conditional *absolute* error variance for a single-facet design in GT. Lord's estimator indeed involves absolute error, a type of error not present in CTT. Therefore, Lord's binomial error model is sometimes considered a bridge between CTT and GT (Brennan, 2010).

Multinomial Error Model

Application of Lord's binomial error model is limited to items with binary score categories (i.e., right or wrong). W. Lee (2007) expanded Lord's model to cases where each item in an n -item test is scored polytomously with k possible score points, $a_1, a_2, \dots, a_k$. A set of n items in a test is viewed as a random sample from an undifferentiated universe of such items. For notational simplicity, the derivation is presented for a single test taker without the subscript p . Let $\eta = \{\eta_1, \eta_2, \dots, \eta_k\}$ indicate the true category-proportion scores for a test taker; namely, η_1 is the proportion of items in the universe for which the test taker would get a score of a_1 , η_2 for a_2 , and so on. The sum of the η values is equal to 1. Let $Y_1, Y_2, \dots, Y_k$ be random variables representing the observed numbers of items scored $a_1, a_2, \dots, a_k$, respectively, such that $Y_1 + Y_2 + \dots + Y_k = n$. It follows that the total raw score X is given as $X = a_1Y_1 + a_2Y_2 + \dots + a_kY_k$.

Due to the random sampling assumption, the category counts $Y_1, Y_2, \dots, Y_k$ will follow a multinomial distribution:

$$f(y_1, y_2, \dots, y_k | \eta) = \frac{n!}{y_1! y_2! \dots y_k!} \eta_1^{y_1} \eta_2^{y_2} \dots \eta_k^{y_k}. \quad (79)$$

The conditional distribution of X can then be obtained as

$$f(x | \eta) = \Pr(X = x | \eta) = \sum_{x=a_1y_1+\dots+a_ky_k} f(y_1, y_2, \dots, y_k | \eta), \quad (80)$$

where the summation is taken over all weighted combinations of $a_1y_1 + \dots + a_ky_k$ that sum to x .

Substituting $\hat{\eta}_i = y_i/n$ for η_i and applying a bias-correction factor, $\sqrt{n/(n-1)}$, an estimator of the CSEM for a test taker with observed category counts of $y_1, \dots, y_k$ is given by

$$CSEM(X)_{MN} = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^k a_i^2 y_i (n - y_i) - 2 \sum_{i < j} a_i a_j y_i y_j \right]}. \quad (81)$$

Of note, using the above formula results in different estimated CSEMs for test takers with the same total raw score but varying configurations of category scores. W. Lee (2007) observed that multinomial CSEMs often display a vertically scattered umbrella pattern when plotted against the total raw score. When item scoring is binary, the multinomial estimates become identical to Lord's estimates.

Univariate GT

A general approach to estimating CSEMs under the GT framework has been developed (Brennan, 1998, 2001c). In GT, a distinction between absolute error and relative error is made even for the CSEMs. Estimating absolute CSEMs is relatively straightforward, while computing relative CSEMs is usually complicated although a simplified formula is available under certain assumptions. In GT, analyses commonly use the *mean* score metric rather than the total score metric, and results expressed in one metric can be easily transformed into the other. As discussed earlier, absolute error for a single person p is defined as $\Delta_p = \bar{X}_p - \mu_p$, the difference between the person's observed mean score and universe score. The variance of Δ_p over randomly parallel instances of a measurement procedure is the absolute error variance for the person. In general, for any random effects design, the absolute CSEM for person p in the *mean* score metric is

$$CSEM(\bar{X})_{GTabs} = \sigma^2(\Delta_p). \quad (82)$$

For example, if the D-study design under consideration is $p \times (I : H)$, the design for an *individual person* is $I : H$. The variance components $\sigma^2(I : H)_p$ and $\sigma^2(H)_p$ are computed using data for each individual person, highlighted by the subscript p below each variance component. Then, the absolute CSEM for person p in the *mean* score metric is $\sigma(\Delta_p) = \sqrt{\sigma^2(h)_p/n'_h + \sigma^2(i : h)_p/n'_i n'_h}$. Multiplying this result by $n'_i n'_h$ gives $\sigma(\Delta_p)$ in the total score metric.

Estimating *relative* CSEMs in GT is much more complicated, largely because of some nonzero covariance terms that are difficult to estimate (see Jarjoura, 1986). For practical purposes, Brennan (2001c) offered an approximate estimator for any random effects designs in the *mean* score metric as

$$CSEM(\bar{X})_{GTrel} = \sigma(\delta_p) \cong \sqrt{\sigma^2(\Delta_p) - [\sigma^2(\Delta) - \sigma^2(\delta)]}. \quad (83)$$

Estimating the relative CSEM using Equation 83 is a two-step process: (a) computing the absolute CSEM using Equation 82 and (b) making an adjustment to it by the difference between the absolute and relative *overall* error variances. Clearly, the adjustment to $\hat{\sigma}^2(\Delta_p)$ is a constant for all persons.

UIRT Proficiency Estimates

In the context of ML estimation, the item and test information functions are frequently used as a measure of conditional accuracy of estimation. If item parameters are known, the information functions do not depend on the group of test takers being tested. The conditional variance of ML estimates is inversely related to the test information function as $\sigma^2(\hat{\theta}|\theta) = 1/I(\theta, \hat{\theta})$ (Lord, 1980). Thus, $CSEM(\hat{\theta}) = \sqrt{1/I(\theta, \hat{\theta})}$. The concept of a test information function and its inverse to define error variance holds numerous convenient properties. A limitation, however, is that there is no general conceptual framework or statistical procedure for incorporating multiple sources of error in the estimation of error variance. An exception is the work by Bock et al. (2002), who proposed an approach to incorporating multiple ratings in IRT aiming for correcting for the bias in the standard error of proficiency estimates. Nevertheless, this approach is considered somewhat ad hoc, lacking a robust substantiation of the notion of replications (e.g., distinction between random and fixed facets) (Brennan, 2001a). Their approach certainly deserves further research.

Error variance in the Bayesian framework is defined as the variance of the posterior distribution of θ for a test taker with item response data $\mathbf{u}$, and its standard deviation is the CSEM for the test taker. Since there is a one-to-one correspondence between the EAP estimator $\tilde{\theta}$ and $\mathbf{u}$, test takers with identical item responses will share the same CSEM. The variance form of CSEM for $\tilde{\theta}$ was presented in Equation 44. Under this framework, the notion of replications plays no explicit role in conceptualizing the variability of θ 's in the posterior distribution. To explicitly incorporate errors due to replications in estimating CSEMs and other reliability statistics, one approach is to apply the estimation procedure to at least two independent administrations of the test on the same group of test takers, preferably using different sets of items from the same domain. While this empirical approach is more feasible with computerized adaptive testing, a comprehensive theoretical framework has not been developed to seamlessly integrate the Bayesian definition of error based on the posterior distribution with one fixed set of data and the errors from actual replications involving more than one data set.

One well-known property of IRT proficiency estimates is that CSEMs tend to be smallest in the middle range of the proficiency scale and increase at both extremes. This phenomenon is not inherent to IRT but is a direct consequence of using the θ scale. As argued by Lord (1980), there is no unique virtue of the θ scale for measuring proficiency, and in general, there is no obvious reason to prefer θ over any other monotonic transformation of it. If θ is transformed to the (true) raw-score metric, the resulting CSEMs will exhibit a reversed pattern compared to that for the θ CSEMs. Nonlinear transformations from one metric to another not only change

the magnitude of CSEMs but also alter their general pattern along the score scale, as discussed next.

Scale Scores

The current *Standards* (AERA et al., 2014) and its predecessors recommend (a) reporting standard errors at different score levels (i.e., conditional) and (b) expressing them in units of a reported score scale. Most reliability statistics, including CSEMs, rely on the specific score scale used for reporting. Two main approaches exist for estimating CSEMs for nonlinearly transformed scale scores. The first utilizes the delta method, a general statistical technique for estimating the variance of a statistic that is a function of another statistic with a known variance (Kendall & Stuart, 1977). The second approach involves computing the conditional distribution for scale scores under assumed psychometric models.

The Delta Method Approximation

To demonstrate how scale transformation affects CSEMs, we begin with the discussion of linear transformation. Let $S = t(X)$ be a linear transformation function of scores X to scale scores S such that $S = B + A(X)$, where A and B are the slope and intercept of the linear function, respectively. It follows that $\sigma^2(S) = A^2 \sigma^2(X)$. This relationship also holds for conditional observed score (i.e., error) variances. That is, denoting E_X and E_S as error scores on the metrics of X and S , respectively, $\sigma^2(E_S | \tau) = A^2 \sigma^2(E_X | \tau)$. This suggests that the scale-score CSEM can be expressed by multiplying the CSEM for scores X by the slope of the transformation function. Thus, for linear transformations, computations are straightforward if CSEM estimates for the original score scale, X , are available.

When dealing with nonlinear transformations, where the slope of the transformation is not constant across score levels, complexities arise. The delta method provides a useful approximation to address this issue. Consider a continuous, differentiable function, $S = t(X)$, where X can represent raw scores, IRT proficiency estimates, or composite scores. The variance of S can be approximated using the delta method as

$$\sigma^2(S) = \left(\frac{dt}{dX}(\mathcal{E}X) \right)^2 \sigma^2(X), \quad (84)$$

where $dt/dX = t'(X)$ is the first derivative function of $t(X)$ with respect to X , which is evaluated at the mean of X . In essence, the delta method approximates the variance of scale scores as a function of both the variance of X and the first derivative (i.e., slope) of the transformation function.

In real-world applications, $S = t(X)$ is rarely a continuous function with a concrete mathematical expression. More frequently, a conversion table contains a discrete set of values for X and corresponding scale-score points, usually rounded to integers. To attain a differentiable mathematical function, Feldt and Qualls (1998) proposed fitting a high-degree polynomial function to the discrete conversion table as

$$t(X) \cong \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q, \quad (85)$$

where the polynomial degree q can be determined based on visual inspection of the fit, incremental R -square statistics, or other relevant criteria. The first derivative of the polynomial function at $X = x$ is

$$t'(x) \cong \beta_1 + 2\beta_2 x + \cdots + q\beta_q x^{q-1}. \quad (86)$$

Once the first derivative function is computed using Equation 86, the scale-score CSEM can be approximated using any CSEM estimators for X . In general, the scale-score CSEM is estimated using this polynomial procedure as

$$CSEM(S)_{PN} = \sigma(E_S | \tau) = \sigma(S | \tau) \cong t'(\mathcal{E}X | \tau) \sigma(X | \tau). \quad (87)$$

For example, consider a nonlinear transformation of IRT θ to a certain type of scale scores with scoring performed using the EAP estimator. The first step is to fit a polynomial regression on the conversion relationship and find the first derivative function as in Equation 86. Then, an approximate value of the scale-score CSEM for a test taker with $\tilde{\theta}$ is $\sigma(S | \tilde{\theta}) \cong t'(\tilde{\theta}) \sigma(\theta | \tilde{\theta})$ because $E(\theta | \tilde{\theta}) = \tilde{\theta}$. The polynomial procedure, as an approximation to the delta method, is versatile because it can be combined with any CSEM estimates for the original untransformed scores. However, a limitation is the somewhat subjective decision on the degree of the polynomial, and the estimated conditional slope based on the fitted function is still an approximation. The subsequent estimation procedures discussed do not necessitate such a slope-estimation process and utilize the entire distribution for scale scores.

Strong True Score Models

Kolen et al. (1992) extended Lord's (1965) strong true score theory to estimating scale-score CSEMs for tests scored number correct. Letting π denote the proportion-correct true score, the general expression for strong true score theory is given by

$$f(x) = \Pr(X = x) = \int_{\pi} \Pr(X = x | \pi) f(\pi) d\pi, \quad (88)$$

where the conditional distribution, $f(x | \pi) = \Pr(X = x | \pi)$, can be modeled using either a binomial or a two-term approximation to the compound binomial distribution; the marginal true score distribution, $f(\pi)$, is assumed to follow either a two-parameter or a four-parameter beta distribution; and $f(x)$ is the resulting marginal observed score distribution. Kolen et al. (1992) considered the most complex four-parameter beta compound binomial model.

The CSEM on the raw-score metric is expressed as

$$CSEM(X)_{ST} = \sigma(X | \pi) = \sqrt{\sum_{x=0}^n x^2 \Pr(X = x | \pi) - \left[\sum_{x=0}^n x \Pr(X = x | \pi) \right]^2}. \quad (89)$$

For a given discrete raw-to-scale score conversion table, the probability for earning a particular raw score is identical to the probability for earning the corresponding scale score—that is, $f[t(x)|\pi] = f(x|\pi)$. This allows replacing X with $S = t(X)$ in Equation 89 to obtain the scale-score CSEM:

$$CSEM(S)_{ST} = \sigma(S|\pi) = \sqrt{\sum_{x=0}^n [t(x)]^2 \Pr(X = x | \pi) - \left[\sum_{x=0}^n t(x) \Pr(X = x | \pi) \right]^2}. \quad (90)$$

Equation 90 provides unambiguous estimates of scale-score CSEMs no matter whether a conversion table is one-to-one or many-to-one (i.e., multiple raw-score points convert to a single scale-score point). If desired, the conditional scale-score distribution based on a many-to-one conversion table can be attained by summing the probabilities associated with all raw scores converting to a particular scale-score point. Doing this for all unique scale-score values will give a distribution for scale scores.

Binomial Procedure

For a test with dichotomous items, Brennan and Lee (1999) derived results for scale-score CSEMs, seen as a scale-score analogue of Lord's CSEM. The assumptions underlying this approach align with Lord's binomial error model. Since the binomial error model does not assume a specific true score distribution, the CSEM is estimated for each individual or uses observed score as a conditioning variable.

The conditional observed score distribution for an individual with a proportion-correct observed score $\bar{x} = \hat{\pi}$ follows a binomial distribution as

$$\Pr(X = x | \hat{\pi}) = \binom{n}{x} \hat{\pi}^x (1 - \hat{\pi})^{n-x}. \quad (91)$$

Lord's CSEM is the square root of the unbiased estimate of variance of the binomial distribution, as shown in Equation 78: $\sqrt{n/(n-1)} \sqrt{[x(n-x)]/n}$, where $\sqrt{n/(n-1)}$ is a bias-correction factor. Lord's CSEM formula can be re-expressed as

$$\sigma(X | \hat{\pi}) = \sqrt{\frac{n}{n-1}} \sqrt{\sum_{x=0}^n x^2 \Pr(X = x | \hat{\pi}) - \left[\sum_{x=0}^n x \Pr(X = x | \hat{\pi}) \right]^2}, \quad (92)$$

where the squared term inside the square root is the mean of the conditional distribution. By replacing raw score x with $t(x)$, the scale-score version of Lord's CSEM is

$$CSEM(S)_{BN} = \sqrt{\frac{n}{n-1}} \sqrt{\sum_{x=0}^n [t(x)]^2 \Pr(X = x | \hat{\pi}) - \left[\sum_{x=0}^n t(x) \Pr(X = x | \hat{\pi}) \right]^2}. \quad (93)$$

Brennan and Lee (1999) discovered that the overall pattern and magnitude of scale-score CSEMs estimated based on the binomial procedure closely aligned with results from the polynomial-delta method.

Other Procedures

The expression for scale-score CSEMs in UIRT can be obtained by the square root of Equation 57. The derivation of scale-score CSEMs for the multinomial error model is outlined by W. Lee (2007), which is virtually the same as the process used for the binomial procedure. It involves using the fact that $f[t(x)|\eta] = f(x|\eta)$ (see Equation 80) and computing the standard deviation of the scale-score distribution.

Composite Scores

Compound Binomial and Compound Multinomial Error Models

The binomial error model has been extended to cases in which a test consists of multiple strata or components, known as the compound binomial model (Feldt, 1984). Let the number of items in each of M components be $n_1, n_2, \dots, n_M$ and $X_1, X_2, \dots, X_M$ represent the component scores. In this model, it is assumed that each item belongs to one of M distinct universes of items, errors follow a binomial distribution within each component, and errors are uncorrelated across components. Due to the uncorrelated-error assumption, the composite error variance is a simple weighted sum of Lord's (1957) error variances for the components. For a test taker with observed component scores $x_1, x_2, \dots, x_M$, the estimated CSEM (often called Feldt's SEM) is

$$CSEM(Z)_{CB} = \sqrt{\sum_{m=1}^M w_m^2 \left[\frac{x_m(n_m - x_m)}{n_m - 1} \right]}. \quad (94)$$

Similarly, the extension of the multinomial error model to composite scores is provided by W. Lee (2007). The composite CSEM is the square root of a weighted sum of multinomial error variances for the various components, each of which is estimated using Equation 81. A constraint for the compound multinomial model is that the number of score categories must be the same for all items within a component.

Multivariate GT

As an extension of univariate GT, multivariate GT provides a framework for estimating CSEMs for composite scores. From the multivariate GT perspective, the foregoing treatment of composite scores is nothing more than a multivariate $p^\bullet \times i^\circ$ design. Let us define the composite score in the *mean* score metric as $\bar{Z} = \sum_{m=1}^M w_m \bar{X}_m$. Under the multivariate $p^\bullet \times i^\circ$ design, the within-person design for each person is simply i° ; that is, items are nested within a set of fixed content categories. In such a case, the absolute CSEM for a person is the square root of a weighted sum of squared absolute CSEMs over the content categories: $\sigma(\Delta_{p\bar{Z}}) = \sqrt{\sum_m w_m^2 \sigma_m^2(\Delta_p)} = \sqrt{\sum_m w_m^2 \sigma_m^2(i) / n'_{im}}$. This result, when transformed to the total score metric, will be identical to the result based on the compound binomial model presented in Equation 94. Similarly, the relative composite score CSEM for a person is $\sigma(\delta_{p\bar{Z}}) = \sqrt{\sum_m w_m^2 \sigma_m^2(\delta_p)}$, where the relative CSEM for each category is computed using Equation 83.

The application of multivariate GT in the estimation of composite score CSEMs is not limited to the $p^\bullet \times i^\circ$ design and can be further extended to many multivariate designs that other competing approaches may not be able to handle properly. For example, consider an assessment that consists of two components, speaking and listening, each consisting of different performance tasks (i), and test takers' responses to both types of tasks are evaluated by the same set of human raters (r). Here, the two content components act as the fixed multivariate variable and the multivariate design is $p^\bullet \times i^\circ \times r^\bullet$. The within-person design is $i^\circ \times r^\bullet$ and the absolute error variance–covariance matrix

is $\begin{bmatrix} \sigma_1^2(\Delta_p) & \sigma_{12}(\Delta_p) \\ \sigma_{12}(\Delta_p) & \sigma_2^2(\Delta_p) \end{bmatrix}$, where the subscripts 1 and 2 indicate the two components,

and the variances and covariances are derived from the variance–covariance component matrices of Σ_i , Σ_r , and Σ_{ir} . Note that errors are allowed to be correlated here, which is unique to multivariate GT compared to other approaches that often assume uncorrelated errors.

A general formula for computing absolute error composite score CSEMs is given by

$$CSEM(\bar{Z})_{MGTab_s} = \sigma_{\bar{Z}}(\Delta_p) = \sqrt{\sum_m w_m^2 \sigma_m^2(\Delta_p) + \sum_{m \neq m'} w_m w_{m'} \sigma_{mm'}(\Delta_p)}. \quad (95)$$

Similar to Equation 83 for the univariate case, the relative error composite score CSEMs can be approximated as

$$CSEM(\bar{Z})_{MGTr_{el}} = \sigma_{\bar{Z}}(\delta_p) \cong \sqrt{\sigma_Z^2(\Delta_p) - [\sigma_Z^2(\Delta) - \sigma_Z^2(\delta)]}, \quad (96)$$

where the correction to $\sigma_Z^2(\Delta_p)$ is the difference between the overall absolute and relative error variances for the composite scores.

MIRT

As previously discussed, the conditional composite score distribution, $\Pr(Z = z | \theta)$, can be obtained using Equation 60 under either the UIRT or the MIRT framework. The standard deviation of the conditional distribution is the CSEM for the composite, which is equal to the square root of Equation 62. For graphical representation of composite-score CSEMs, especially under MIRT, it is common to use the true composite score, ζ_θ , as the conditioning variable. However, the relationship between θ and ζ_θ in MIRT is not strictly one-to-one, leading to challenges in visualization. This is because of the infinite possible combinations of component θ 's that yield the same composite true score. Thus, multiple CSEM estimates can be associated with each true composite score. To address this issue, various strategies can be employed, such as computing the arithmetic mean for each true composite score or applying a polynomial regression to obtain single-valued smoothed estimates.

A MIRT model typically provides better fit than a UIRT model when a test is composed of multiple subdimensions such as content areas. Oftentimes, however, a UIRT model is judged to be robust to a moderate violation of the unidimensionality assumption and more feasible to use in practice because of its computational simplicity and unambiguous interpretation of results. When applied to mixed-format tests, W. Lee et al. (2020) observed that the UIRT and MIRT models generally agreed closely in terms of the patterns of CSEMs; however, when aggregated to obtain the overall SEM and reliability, results tended to be quite different as the data became more multidimensional.

Summary and Other Issues

The procedures discussed in this section are summarized in Table 5.3, categorized based on their respective model, score types, assumptions about replications, and item format. Not covered in the preceding section are methods for estimating CSEMs for scale scores converted from composite scores. Composite scale-score CSEMs can be derived using the compound binomial model (Brennan & Lee, 1999), compound multinomial model (W. Lee, 2007), or MIRT (W. Lee et al., 2020).

When raw scores or IRT proficiency estimates are the primary focus, estimating CSEMs is relatively straightforward, and the general shapes of these CSEMs are highly predictable. However, in many cases, raw scores or IRT proficiency estimates are transformed into scale scores for reporting. In such cases, expressing CSEMs on the metric of a reported score scale provides more informative results. Previous research suggests that the overall pattern of scale-score CSEMs along the score scale tends to follow the changes in the slope (or first derivative) of a transformation function (Brennan & Lee, 1999; Feldt & Qualls, 1998; Kolen & Lee, 2011; W. Lee et al., 2000, 2020). A steeper slope at a raw-score point corresponds to a wider range of adjacent raw scores converting to a broader range of scale scores, leading to greater variability of scale scores over repeated measurements. Conversely, a flatter slope is associated with a wide range of raw scores converting to a smaller range of scale scores, resulting in less variability of scale scores over replications. It is advisable in practice that the relative magnitude of scale-score CSEMs be examined, especially near the score points that are critical for decision-making, such as cut scores for licensure and certification exams.

The delta method approximation can serve as a general procedure for estimating scale-score CSEMs because it can be used jointly with almost any estimates of raw-score CSEMs. In most applications in educational testing contexts, the delta method requires an approximation of the first derivative function through a *fitted* conversion table to obtain a continuous, differentiable function. A high-degree polynomial model is often chosen due to its flexibility and ease of differentiation. However, the performance of the delta method with polynomial approximation may be contingent on the characteristics of a conversion table. For example, if a conversion table is characterized as many-to-one, the polynomial approximation of slope may

Table 5.3 Estimators of Conditional Standard Errors of Measurement Based on a Single Test Administration

Procedure/Model	Score Type	Item Type	Assumption
Thorndike	Raw score	Both	Essential tau-equivalent split halves
Mollenkopf	Raw score	Both	Essential tau-equivalent split halves
Binomial error model	Raw score Scale score	DI	Randomly parallel
Compound binomial error model	Composite score Composite scale score	DI	Stratified randomly parallel
Multinomial error model	Raw score Scale score	PO	Randomly parallel
Compound multinomial error model	Composite score Composite scale score	PO	Stratified randomly parallel
Univariate GT	Raw score: Absolute error Relative error	Both	Randomly parallel
Multivariate GT	Composite score: Absolute error Relative error	Both	Stratified randomly parallel
Strong true score model	Raw score Scale score	DI	Randomly parallel
Unidimensional IRT	Raw score Scale score ML estimator	Both	Strictly parallel
Unidimensional IRT	EAP estimator	Both	Undefined
Multidimensional IRT	Composite score Composite scale score	Both	Stratified strictly parallel
Delta-polynomial method	Scale score Composite scale score	Both	

Note. DI = dichotomous item type; PO = polytomous item type.

be questionable. In that case, estimation procedures that leverage the conditional scale-score distribution may be deemed preferable, although further research is needed to substantiate this preference.

Virtually all estimation procedures for scale-score CSEMs discussed in this section assume a *constant* raw-to-scale score conversion table across hypothetically replicated parallel forms and for every test taker in the population. In real-world testing situations where multiple alternate forms of a test are developed and used, the concern is centered around the interchangeability of scores across these forms, which may vary in difficulty. Equating plays an important role in this context, and an

outcome of such equating is a raw-to-scale score conversion table *specific* to each test form. The resulting scale scores are considered interchangeable across different forms if equating is trustworthy.

The need for equating to generate different conversion tables for each hypothetical replication in estimating scale-score CSEMs has not yet been systematically investigated in the literature. Obviously, the assumption of strictly parallel forms for the IRT procedures does not require equating because forms are effectively fixed, and any score differences are solely attributable to random error. In contrast, the assumption of randomly parallel forms in binomial and multinomial procedures, at least in the context of estimating CSEMs, does not require every test taker to take the same set of items on each replication. This implies that a test form, as a collection of specific items, is not clearly defined under this conception. If an additional restriction is imposed to allow for the same random sample of items to be used for all test takers in each replication, the resulting forms will have equivalent difficulty levels because all items selected for each replication are of equal difficulty for any given test taker. Therefore, under the randomly parallel forms assumption, equating is deemed either impossible or unnecessary. Further exploration of this issue may be warranted.

While the overall and conditional SEMs by themselves provide useful information about the amount of error in test scores, they are frequently used in conjunction with intervals. The *Standards* (AERA et al., 2014), especially Standard 6.10, recommend that score precision be depicted by error bands using the SEM. There is a voluminous literature on various types of interval estimation procedures in relation to different types of scores. Space does not permit any systematic review of the topic, but the following two principles merit consideration. First, the CSEM is usually favored over the overall SEM when constructing intervals for each individual test taker. Second, the endpoints of an interval should preferably be on the score metric used for reporting (e.g., scale scores). For example, W. Lee et al. (2006) recommended using CSEMs rather than the overall SEM based on a simulation study. They also proposed an endpoints conversion method for constructing intervals for nonlinearly transformed scale scores from raw scores, demonstrating superior performance compared to a normal approximation method.

As a final note, if a test has a multidimensional structure by design, such as a table of content specifications or mixed item formats, it is strongly recommended that the CSEMs for the total scores be estimated using one of the composite score models. Ignoring the fact that domains are stratified and fitting noncomposite models to the total scores could introduce bias in the estimated CSEMs, with larger bias as data become more multidimensional (W. Lee et al., 2000).

RELIABILITY OF CLASSIFICATION CATEGORY SCORES

When using a test score or composite score for categorical classification decisions (e.g., pass/fail), the *Standards* (AERA et al., 2014) recommend estimating the consistency of classifications across two replications of the same measurement procedure.

Classification errors occur when decisions are based on test takers' observed scores, which contain measurement error. In this context, the interpretation of test scores primarily considers each test taker's observed score relative to a standard or cut score, rather than the performance of other test takers. Since Glaser (1963) introduced the notion of criterion-referenced interpretation of test scores, extensive research has been devoted to developments of indices for measuring the precision of classifications. Berk (1980) provided a comprehensive summary of the measures developed until 1980, and since then, new methods have emerged as a result of increased use of IRT and complex assessment types.

This section reviews procedures based on threshold-error loss for assessing consistency of classifications. These methods primarily focus on the extent of consistent classifications with respect to the cut score, treating all false classifications as equally serious. By contrast, squared-error loss methods, such as those proposed by Livingston (1972), Brennan and Kane (1977a, 1977b), and Kane and Brennan (1980), consider misclassifying test takers with scores far from the cut score as more serious than misclassifying those with scores near the cut score. While both approaches offer meaningful insights into classification consistency, they address different aspects of the classification problem. Another group of procedures, not reviewed in this section, includes methods involving dividing the test into two parallel halves and stepping up the result to what might be expected for the full-length test (e.g., Breyer & Lewis, 1994; Woodruff & Sawyer, 1989).

The precision, or lack thereof, of classifications is often described in terms of classification consistency and accuracy in the literature. *Classification consistency* measures the degree to which test takers are classified in the same performance category based on two independent replications of the same (or similar) measurement procedure. A double administration procedure (Hambleton & Novick, 1973) using two parallel forms involves tallying the proportion of test takers assigned to each classification category on both administrations. By contrast, single-administration procedures rely on two expected (i.e., model-predicted) observed score distributions from two hypothetical replications of the test. The concept of replications, whether actual or hypothetical, leads to the perception of classification consistency as the reliability of classifications (W. Lee et al., 2002).

Classification accuracy assesses the agreement between classifications based on test takers' observed scores and classifications based on test takers' true scores (W. Lee et al., 2002; Livingston & Lewis, 1995). Unlike classification consistency, which compares two observed classifications, classification accuracy relates observed classifications to true classifications. In this sense, classification accuracy is sometimes referred to as decision validity (Berk, 1980; Hambleton, 1980).

Classification Consistency and Accuracy Indices

Suppose that test takers are classified into H mutually exclusive performance categories based on a set of $H - 1$ cut scores on the raw-score metric, $c_1, c_2, \dots, c_{H-1}$. For notational

convenience, let c_0 and c_H denote the lowest and highest possible scores, respectively. The first performance category contains scores ranging from c_0 to $c_1 - 1$; the second category contains scores from c_1 to $c_2 - 1$; and so forth. Let $P_1, P_2, \dots, P_H$ denote the observed performance categories in which each test taker is classified by comparison of their observed scores to cut scores. Further, let X_1 and X_2 be the random variables for observed scores on two administrations of the test that are independent and identically distributed.

The two most frequently discussed classification consistency indices are the agreement index ϕ and Cohen's (1960) kappa (κ) coefficient. The marginal agreement index ϕ is the percentage of test takers consistently classified in the same category on two independent replications:

$$\phi = \sum_{h=1}^H \Pr(X_1 \in P_h, X_2 \in P_h). \quad (97)$$

More specifically, the ϕ coefficient is the sum of diagonal elements of an $H \times H$ contingency table composed of joint probabilities or percentages of observed category classifications on two replications, denoted $\Pr(X_1 \in P_i, X_2 \in P_j)$. The κ coefficient adjusts for agreement occurring by chance and is given by

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}, \quad (98)$$

where ϕ_c is the chance agreement: $\phi_c = \sum_{h=1}^H \Pr(X_1 \in P_h) \Pr(X_2 \in P_h) = \sum_{h=1}^H [\Pr(X_1 \in P_h)]^2$ because X_1 and X_2 are identically distributed.

Classification accuracy relies on a bivariate distribution of observed and true classifications. True classification involves determining a test taker's true categorical status based on their true score relative to cut scores expressed on the true score metric, $\lambda_1, \lambda_2, \dots, \lambda_{H-1}$, with λ_0 and λ_H being the lowest and highest possible true scores, respectively. These true cut scores typically are set to be the same as the observed cut scores, but theory does not require doing so. Because test takers' true scores are unknown, individual true scores can be estimated, or an entire true score distribution for a population can be estimated or assumed. Let $\Gamma_h (h = 1, 2, \dots, H)$ denote the true category of a test taker with true score τ such that $\lambda_{h-1} \leq \tau < \lambda_h$. An $H \times H$ contingency table can be generated, containing the joint probabilities of observed and true classifications, $\Pr(\tau \in \Gamma_i, X \in P_j)$. This contingency table is asymmetric. The marginal classification accuracy index, denoted γ , is defined as the sum of the diagonal elements in the contingency table:

$$\gamma = \sum_{h=1}^H \Pr(\tau \in \Gamma_h, X \in P_h). \quad (99)$$

If the rows represent true categories and columns represent observed categories, the sum of the upper diagonal elements indicates the percentage of test takers classified in observed categories higher than their true category, known as the marginal false-positive error rate, symbolically expressed as

$$\gamma^+ = \sum_{i=1}^{H-1} \sum_{j=i+1}^H \Pr(\tau \in \Gamma_i, X \in P_j). \quad (100)$$

Conversely, the marginal false-negative error rate represents the percentage of test takers whose observed categories are lower than their true categories and is defined as the sum of the lower diagonal elements of the contingency table:

$$\gamma^- = \sum_{i=2}^H \sum_{j=1}^{i-1} \Pr(\tau \in \Gamma_i, X \in P_j). \quad (101)$$

In essence, all estimation procedures discussed next involve, one way or another, constructing $H \times H$ contingency tables for consistency and accuracy. Nevertheless, they differ with respect to underlying assumptions ascribed to various models and applicable item types (dichotomous vs. polytomous) and score types (raw, composite, or theta estimates).

For most estimation methods, it is more convenient and informative to begin with estimating classification indices for each individual test taker or conditional on each level of true score. These conditional indices are then aggregated or averaged to derive the marginal indices. The subsequent sections primarily concentrate on the ϕ and γ coefficients to simplify matters.

Strong True Score Models

The first procedure is grounded in strong true score models. Huynh (1976) introduced an estimation procedure based on the beta-binomial model, the simplest form of a family of strong true score models. Later, Hanson and Brennan (1990) extended Huynh's approach to more general strong true score models. The conditional raw-score distribution, $f(x|\pi)$, can be modeled using either a binomial or a two-term approximation to the compound binomial model. The distribution of π for a population is either a two-parameter beta or a four-parameter beta distribution.

Let $f(x_1|\pi)$ and $f(x_2|\pi)$ represent the identically distributed conditional raw-score distributions on two replications of a measurement procedure. The conditional probability of being classified in the h th observed category P_h is

$$\Pr(X_1 \in P_h | \pi) = \Pr(X_2 \in P_h | \pi) = \sum_{x_1=c_{h-1}}^{c_h-1} f(x_1|\pi). \quad (102)$$

Due to the local independence assumption, the conditional joint distribution of X_1 and X_2 is

$$f(x_1, x_2 | \pi) = f(x_1 | \pi) f(x_2 | \pi). \quad (103)$$

Given a set of $H - 1$ cut scores, a *conditional* $H \times H$ contingency table can be generated based on the conditional joint probabilities. The probability of falling in the same h th category on two replications (i.e., diagonal elements) can be computed as

$$\Pr(X_1 \in P_h, X_2 \in P_h | \pi) = \sum_{x_1=c_{h-1}}^{c_h-1} f(x_1 | \pi) \sum_{x_2=c_{h-1}}^{c_h-1} f(x_2 | \pi) = \left[\sum_{x_1=c_{h-1}}^{c_h-1} f(x_1 | \pi) \right]^2. \quad (104)$$

It follows that the conditional probability of consistent classifications is

$$\phi(\pi) = \sum_{h=1}^H \Pr(X_1 \in P_h, X_2 \in P_h | \pi), \quad (105)$$

which is referred to as the conditional agreement index. The marginal agreement index is obtained by integrating $\phi(\pi)$ over the entire distribution of π :

$$\phi = \int_{\pi} \phi(\pi) f(\pi) d\pi, \quad (106)$$

where integration is performed numerically using a set of quadrature points from an estimated true score distribution (e.g., beta).

Now, let us suppose that a true score (or a quadrature point) belongs to the h th performance level, $\pi \in \Gamma_h$. An accurate classification occurs when, based on observed scores, a test taker is classified in a performance category that is the same as the test taker's true performance category. Therefore, the conditional accuracy index is simply

$$\gamma(\pi) = \Pr(X \in P_h | \pi \in \Gamma_h), \quad (107)$$

which can be computed using Equation 102. The marginal accuracy index is then given by

$$\gamma = \int_{\pi} \gamma(\pi) f(\pi) d\pi. \quad (108)$$

The conditional and marginal false-positive and -negative error rates can be derived in a similar manner.

Normal Approximation Procedure

The normal approximation procedure (Peng & Subkoviak, 1980) was initially proposed as a simplification of Huynh's (1976) beta-binomial procedure. Research suggests, however, that this relatively simple procedure performs adequately across various situations and can accommodate different types of scores (e.g., S. Y. Kim & Lee, 2019, 2020; Wan et al., 2007). The crucial assumption of this procedure is that the observed scores from two replications follow a bivariate normal distribution with a correlation equal to test reliability.

Let $z(c_h)$ denote a z -score corresponding to each cut score c_h ($h = 1, 2, \dots, H$). This standardization enables us to use the convenient properties of z scores. The marginal

percentages of test takers who are classified in the same h th performance category on two replications, $\Pr(X_1 \in P_h, X_2 \in P_h)$, can be determined using a cumulative standard bivariate normal distribution. The marginal agreement index ϕ is a sum of $\Pr(X_1 \in P_h, X_2 \in P_h)$ over all performance categories, as follows:

$$\begin{aligned}\phi &= \sum_{h=1}^H \Pr(X_1 \in P_h, X_2 \in P_h) \\ &= \sum_{h=1}^H \int_{z(c_{h-1})}^{z(c_h)} \int_{z(c_{h-1})}^{z(c_h)} \frac{1}{2\pi\sqrt{1-R(X)^2}} \exp\left[-\frac{x_1^2 - 2R(X)x_1x_2 + x_2^2}{2[1-R(X)^2]}\right] dx_1 dx_2, \quad (109)\end{aligned}$$

where $R(X)$ is an estimate of reliability. Based on the fact that the marginal distribution of X_1 (or X_2) is univariate normal, the marginal category probabilities of $\Pr(X_1 \in P_h)$ and $\Pr(X_2 \in P_h)$ can be computed using a cumulative univariate standard normal distribution. The chance agreement and kappa coefficient can be computed accordingly.

Classification accuracy was not considered explicitly by Peng and Subkoviak (1980). Later researchers adopted the normality-based framework for estimating classification accuracy indices (e.g., S. Y. Kim & Lee, 2019). The assumption is that the marginal joint distribution of the observed and true scores follows a bivariate normal distribution. Since both observed and true classifications are involved, the true cut scores as well as the observed cut scores need to be standardized. For summed raw scores, $\mu(X) = \mu(T)$ and $\sigma(T) = \sigma(X)\sqrt{R(X)}$ in CTT. Thus, true cut scores are standardized as $z(\lambda_h) = [\lambda_h - \mu(X)] / \sigma(X)\sqrt{R(X)}$. The correlation of this bivariate distribution is $\sqrt{R(X)}$ because $\rho(X, T) = \sqrt{R(X)}$. The marginal accuracy index is given by

$$\begin{aligned}\gamma &= \sum_{h=1}^H \Pr(X \in P_h \mid \tau \in \Gamma_h) \\ &= \sum_{h=1}^H \int_{z(c_{h-1})}^{z(c_h)} \int_{z(\lambda_{h-1})}^{z(\lambda_h)} \frac{1}{2\pi\sqrt{1-R(X)}} \exp\left[-\frac{\tau^2 - 2\sqrt{R(X)}\tau x + x^2}{2[1-R(X)]}\right] d\tau dx. \quad (110)\end{aligned}$$

A cumulative univariate standard normal distribution for each of the true and observed scores can be used to obtain the marginal category probabilities of $\Pr(\tau \in \Gamma_h)$ and $\Pr(X \in P_h)$, which then are used to fill the off-diagonal cells of the marginal $H \times H$ contingency table, $\Pr(\tau \in \Gamma_i, X \in P_j)$, for computing false-positive and -negative error rates.

The assumption of bivariate normality for the normal approximation method may not fully hold in reality. However, research suggests that this approach is quite robust to violations of the normality assumption (e.g., S. Y. Kim & Lee, 2020; Wan et al., 2007). Regarding the choice of reliability coefficient in Equations 109 and 110, it is generally preferred to use reliability coefficients that involve absolute error variance because the

focus in classification contexts is on the discrepancy between test takers' scores and cut scores rather than their ranking. Peng and Subkoviak (1980) used KR21 with the normal approximation method, and Huynh (1976) also employed KR21 in the beta-binomial model framework, following Lord and Novick (1968), who demonstrated that reliability for the beta-binomial model is indeed KR21. Although the choice of reliability coefficient may have minimal practical impact, it is prudent to select one that is defensible for a particular application, especially paying close attention to the score type of interest.

Livingston and Lewis Procedure

The strong true score models (Hanson & Brennan, 1990; Lord, 1965) assume that n items are dichotomously scored and equally weighted. The Livingston and Lewis (1995) procedure is intended to extend the strong true score models to other complex types of test scores such as scores from polytomous items, composite scores, and even scale scores. For this somewhat ad hoc extension, they introduced the concept of *effective test length*, denoted here as $\tilde{n}$, which represents the pseudo number of equally difficult dichotomous items required to achieve the same reliability as the actual reported scores. (Note that the term effective test length carries a different interpretation compared to its usage in Equation 9). Letting X denote the actual test scores, the effective test length is computed as

$$\tilde{n} = \frac{[\hat{\mu}(X) - \min(X)][\max(X) - \hat{\mu}(X)] - R(X)\hat{\sigma}^2(X)}{\hat{\sigma}^2(X)[1 - R(X)]}, \quad (111)$$

where $R(X)$ is a reliability coefficient for the original score type, and $\min(X)$ and $\max(X)$ indicate the minimum and maximum possible scores, respectively. Since $\tilde{n}$ can be a noninteger value, it is often rounded to the nearest integer. One of the strong true score models can be applied to this pseudo dichotomous data to compute classification indices as outlined in the section "Strong True Score Models."

The flexibility of the Livingston and Lewis procedure makes it one of the most frequently cited and used procedures in practice. Many comparison studies have found that results for the Livingston and Lewis procedure are comparable to those for other methodologies (e.g., S. Y. Kim & Lee, 2020; W. Lee et al., 2009; Wan et al., 2007).

Binomial and Multinomial Models

Subkoviak (1976) proposed a procedure for estimating classification consistency for a group of test takers under the assumption of dichotomously scored items. Unlike Huynh (1976), who utilizes the full feature of the beta-binomial model, Subkoviak's procedure does not entail estimating the true score distribution, although it assumes a binomial distribution for the conditional observed score distribution, $f(x|\pi)$. Instead of estimating the entire true score distribution, Subkoviak's procedure uses an estimated true score for each individual test taker. Subkoviak (1976) suggests using either a test

taker's observed proportion-correct score or Kelley's (1947) regressed score estimate. The procedure computes the conditional agreement index for each person and then obtains the marginal agreement index by averaging the conditional results over all persons in the data set.

W. Lee (2005) and W. Lee et al. (2009) extended Subkoviak's procedure by: (a) proposing a procedure based on the multinomial error model (W. Lee, 2007) for polytomous items; (b) applying the compound multinomial model (W. Lee, 2007) to composite scores; and (c) considering classification accuracy, which was not addressed in Subkoviak (1976). The multinomial procedure, briefly introduced below, reduces to Subkoviak's model if items are dichotomously scored.

As previously defined in the section "Multinomial Error Model", for a person with $\eta = \{\eta_1, \eta_2, \dots, \eta_k\}$, the random variables, $Y_1, Y_2, \dots, Y_k$, represent the numbers of items scored $a_1, a_2, \dots, a_k$, respectively, which follow a multinomial distribution. The total raw score is $X = \sum_{i=1}^k a_i Y_i$. The conditional total score distribution given η is provided in Equation 80. The probability of being in the h th performance level is given by

$$\Pr(X \in P_h | \eta) = \sum_{x=c_{h-1}}^{c_h-1} f(x|\eta). \quad (112)$$

For a person with $\hat{\eta}_i = y_i/n$, the conditional agreement index, $\phi(\hat{\eta})_p$, is computed using Equations 104 and 105. These conditional agreement indices for all N persons are averaged to obtain the marginal agreement index:

$$\phi = \sum_{p=1}^N \phi(\hat{\eta})_p. \quad (113)$$

Using observed score x as an estimate of true score, each person's true performance level is determined by comparing x to true cut scores. For a person with observed $\hat{\eta}$ and $\lambda_{h-1} \leq \hat{\tau} = x = \sum_{i=1}^k a_i y_i < \lambda_h$, the probability of accurate classifications is simply the probability of being classified into the h th performance category:

$$\gamma(\hat{\eta})_p = \Pr(X \in P_h | \hat{\tau} \in \Gamma_h), \quad (114)$$

which is computed using $\hat{\eta}$ in place of η in Equation 112. The gamma coefficient for a group of test takers is

$$\gamma = \sum_{p=1}^N \gamma(\hat{\eta})_p. \quad (115)$$

The methodology discussed above can be extended to cases where the score type of interest is a composite score. The key difference is to model errors according to the compound binomial model, as discussed in detail by W. Lee (2005, 2007) and W. Lee

et al. (2009). Brennan and Lee (2008) considered correcting for bias in classification consistency indices by proposing an estimator of true score that has the same variance as true scores.

IRT Procedures

Several procedures exist for computing classification indices under the IRT framework. For example, Huynh (1990) considered the Rasch model for classification consistency, with subsequent extensions to accommodate other UIRT models (W. Lee, 2010; W. Lee et al., 2002; Schulz et al., 1999; T. Wang et al., 2000). S. Y. Kim and Lee (2019) and W. Lee et al. (2020) further expanded the UIRT procedures to handle composite scores using assumptions of SS-MIRT and BF-MIRT models. While these procedures were primarily designed for tests based on summed scoring, those developed by Rudner (2001, 2005) and Guo (2006) are geared toward ML proficiency estimates. We first review Rudner's proficiency-based procedure, followed by methods tailored for summed scoring.

ML Proficiency Estimates

Rudner's procedure operates under the assumption that, for person p with an ML estimate $\hat{\theta}_p$, errors are asymptotically normal with a mean of $\hat{\theta}_p$ and standard deviation of $\hat{\sigma}_p = \sqrt{1 / \hat{I}(\hat{\theta}, \hat{\theta})}$, where $\hat{I}(\hat{\theta}, \hat{\theta})$ is a sample estimate of the test information function evaluated at $\hat{\theta}_p$. The expected probability of falling in the h th category is expressed as

$$\Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p) = \Phi(\hat{\theta}_p, \hat{\sigma}_p, c_h) - \Phi(\hat{\theta}_p, \hat{\sigma}_p, c_{h-1}), \quad (116)$$

where $\Phi(\hat{\theta}_p, \hat{\sigma}_p, cut)$ is the cumulative probability below the theta cut score, cut , based on the normal distribution, $N(\hat{\theta}_p, \hat{\sigma}_p)$. Assuming errors are identically distributed for two replications, the probability of consistent classifications for each person is calculated as

$$\phi(\hat{\theta})_p = \sum_{h=1}^H \Pr(\hat{\theta}_1 \in P_h, \hat{\theta}_2 \in P_h \mid \hat{\theta}_p) = \sum_{h=1}^H \left[\Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p) \right]^2. \quad (117)$$

For a test taker with $\lambda_{h-1} \leq \hat{\theta}_p < \lambda_h$, the probability of accurate classifications is equal to the probability of being placed in the h th category:

$$\gamma(\hat{\theta})_p = \Pr(\hat{\theta} \in P_h \mid \hat{\theta}_p \in \Gamma_h). \quad (118)$$

The marginal indices of ϕ and γ for a group of persons are computed by averaging conditional indices as in Equations 113 and 115.

Guo (2006) proposed a modification of Rudner's procedure that relaxes the normality assumption. Instead of using the ML point estimate of θ and its standard error, Guo's approach utilizes the likelihood function of a test taker's item responses

given item parameters to compute probabilities of classifications. Research has shown that both Rudner's and Guo's procedures tend to yield similar results (Guo, 2006; Wyse & Hao, 2012).

Summed Scoring

The formulation of IRT summed-score procedures is virtually the same as, or similar to, that of other procedures, particularly the strong true score models. However, there are two main distinctions: (a) the models used to derive the conditional observed score distribution (i.e., errors) and (b) the conditioning variables. For example, conditional distributions are defined for a UIRT model in Equation 50 or a MIRT model for composite scores in Equation 60. Subsequently, the steps outlined in Equations 102 through 108 can be applied to compute classification indices by replacing π with either θ or θ depending on the model used.

Aggregation Methods

Estimating marginal classification indices requires a distribution of θ (or θ) for a population of test takers. While a discrete quadrature distribution is typically employed for a UIRT model, computational complexity grows exponentially with a MIRT model as the number of dimensions increases. To address this challenge, W. Lee et al. (2020) suggested alternative approaches for aggregating conditional indices. These include using: (a) quadrature distributions (D method), (b) individual proficiency estimates (P method), and (c) Monte Carlo simulation (M method). These three marginalization methods can also be applied to estimating CSEMs and reliability coefficients (see W. Lee et al., 2020).

The D method involves specifying a discrete set of combinations of quadrature points and associated weights for θ , often using a multivariate standard normal distribution. For example, with four dimensions and 21 quadrature points per dimension, there are $21^4 = 194,481$ possible combinations of quadrature points. Conditional agreement indices, $\phi(\theta)$, are computed for each combination and aggregated over the entire θ distribution to obtain the marginal index: $\phi = \sum_q \phi(\theta)w_q$, where the sum is taken over all combinations of theta values and w_q is the weight associated with quadrature combination q .

In contrast, the P method utilizes proficiency estimates as the conditioning variable for each person. Conditional indices are computed for each person with $\hat{\theta}$, and a marginal index is obtained by averaging the conditional indices over the number of persons in the group: $\phi = \sum \phi(\hat{\theta})_p / N$. The P method is typically computationally less intensive than the D method; however, results may be subject to the chosen proficiency estimator. The P method would be most useful when the purpose is to quantify classification errors for each individual test taker.

The M method combines elements of the D and P methods. Similar to the D method, it begins with an assumed distribution of θ , from which a large number of random

deviates are drawn. For example, $N_r = 100,000$ combinations of theta values could be drawn from a multivariate standard normal distribution. Conditional results are estimated for each of the $N_r = 100,000$ simulated test takers. Then, similar to the P method, marginal results are computed by averaging the conditional results over the N_r test takers: $\phi = \sum \phi(\theta) / N_r$.

While these methods typically produce similar results (W. Lee, 2010; W. Lee et al., 2020), the M method is often preferred for MIRT models. It offers computational efficiency, eliminates unrealistic theta combinations, and yields conditional estimates—such as classification errors and SEMs—that are visually more interpretable when plotted.

Other Issues

The foregoing discussion on classification errors is restricted to cut scores established on untransformed score metrics. However, in many large-scale educational testing programs, cut scores are often set on reported score scales. Once a score scale is established for the initial test form, it is maintained across subsequent forms through test equating (Kolen & Brennan, 2014). In this scenario, there are at least two ways to estimate classification indices for scores on each test form. The easiest method would be to use the operational raw-to-scale score conversion table, in which the raw-score points that correspond to the cut scores on the scale-score metric could be identified. Alternatively, the conditional distribution of scale scores can be computed, as discussed in the section “Estimators of CSEMs.” Then, the processes of constructing category probabilities and computing conditional and marginal indices, outlined in the section “Reliability of Classification Category Scores,” can be applied.

The marginal classification indices provide useful information about the overall precision of classifying a group of test takers. However, they do not provide information about the precision of classifications at different score or proficiency levels. For this purpose, conditional indices along the score scale can be examined. Studies have consistently shown that classification errors tend to be larger near cut scores, as one might expect. Examples of exploring conditional classification indices are found in W. Lee (2010), W. Lee et al. (2002, 2020), and Wyse and Hao (2012). Wainer et al. (2005) considered a Bayesian method, involving the Markov Chain Monte Carlo procedure to produce samples from the posterior distribution of a test taker’s proficiency estimate. The probability of passing, what they call the posterior probability of passing curve, is determined by counting the number of sampled proficiency values that exceed the cut score.

Most research to date on classification consistency and accuracy has focused on measurement errors that are attributable to items or forms. It seems natural to consider an extension of the current theories and methodologies to broader measurement situations. For example, classification decisions could be made based on test takers’ responses to a set of essay prompts scored by raters. In such cases, an effective

procedure should allow measurement errors attributable to essay prompts and raters to be modeled separately in constructing performance category probabilities. GT might prove to be a useful tool for developing such estimation models in classification contexts.

OTHER MODELS, AGGREGATION, AND PRECISION ISSUES

There are many additional aspects of reliability this chapter can only touch on.

Other Models and Assessment Types

Diagnostic Classification Models

With diagnostic classification models (DCMs), the latent traits are assumed to be categorical. This implies, typically, that there is a more limited range of possible latent values (e.g., master and nonmaster) than actual score values (raw scores, IRT proficiency estimates, or scale scores). In DCMs, the error variance and the latent trait variance are not independent, causing some complications in estimating reliability.

Using a replication definition of reliability, simulations are used to administer an assessment (the same form, a parallel form, or M. S. Johnson and Sinharay's (2018) definition of multiple assessments with an identical Q matrix and item parameters) repeatedly. What category test takers were assigned to each time was observed. Several authors have investigated different methods of estimating precision in DCMs, including M. S. Johnson and Sinharay (2018), who noted that simply reporting the estimated probability that an individual test taker is correctly classified may be misleading without considering the distribution of the categorical trait in the population. For instance, if 95% of a population are "masters," simply classifying every test taker as a master will result in 100% of the masters being correctly classified and 100% of the nonmasters being incorrectly classified.

Templin and Bradshaw (2013) introduced a measure of DCM reliability and compared it to an IRT measure. Replications for the DCM method can be obtained by repeated sampling from test takers' posterior distributions. For the IRT model estimate, the observed item responses provide an estimate of the latent trait, the standard error of which can be calculated using the test information function. Resampling via simulation can be used to obtain the distribution of the latent variable across the test-taker population, and the correlation computed between scores from two administrations can be used as an IRT reliability estimate. Comparing reliability estimates over the hypothetically repeated administrations, Templin and Bradshaw illustrated both theoretically and through the simulations that DCMs exhibited greater precision than IRT model-based estimates. Thompson et al. (2019) looked at DCM results for an operational assessment system reporting information, including reliability, at five levels and also found that the DCM reliabilities based on

simulations approximating replications of testing were higher than those for CTT or IRT methods.

Structural Equation Models

Structural equation models allow for comparing various models for assessment data, which aids in the determination of an appropriate reliability coefficient in situations where one is uncertain about the dimensionality, tau equivalency, and other aspects of the data. Graham (2006) provided a hierarchy of models to estimate reliability, using structural equation models to test model fit to data, lamenting that

many students and researchers in education and psychology are unaware of many of the assumptions required by a given statistical procedure, are unaware of how to test those assumptions, or are unaware of acceptable alternatives should those assumptions not be met. (p. 942)

Bacon et al. (1995) looked at the underestimation of reliability using coefficient alpha and omega across different conditions, advocating for weighted omega. Hancock and An (2018) provided an introduction to scale reliability within the context of confirmatory factor analysis and structural equation models, endorsing the use of McDonald's ω instead of Cronbach's alpha. S. B. Green and Yang (2008) derived a nonlinear structural equation model method to compute reliability for ordered categorical items. Raykov and Shrout (2002) presented a method to compute point and bootstrapped confidence interval estimates of reliability for weighted and unweighted composites with a general structure. Karimi (2015) illustrated structural equation models in estimating model-based reliability, including applications for the bifactor and mixed reflective-formative models, as well as covariate-dependent and covariate-free reliability.

Raykov and Penev (2010) outlined a latent variable analysis approach to provide both point and confidence interval estimates of reliability of group means, the overall reliability of group means, and conditional reliability for conditional and unconditional two-level models. Kano and Azuma (2003) discussed both the essential tau equivalence and the independence assumptions in a structural equation model environment for coefficient alpha, stating, "the independence assumption is more important than the essential τ equivalence assumption because dependency among unique factors can cause overestimation of the true reliability" (p. 147).

Oberski and Satorra (2013) addressed the "usual practice" of using error variances from an independent study as population variances in subsequent structural equation model analyses, showing how this may make the structural parameters' standard error too small. They provided an adjustment factor, illustrating its use with simulations and empirical data. Structural equation modeling continues to be a productive area of research related to precision coefficients.

Reliability Related to Longitudinal Data and Growth Models

Examining trends, such as an individual's scores over time or pretest and posttest scores to examine the effectiveness of a treatment, involves additional precision issues. Rogosa and Willett (1983) observed that when test takers have similar true growth changes, difference scores cannot adequately distinguish the amounts of true change among test takers. However, if the test takers differ in the amount of true growth, difference scores may indeed be adequate in distinguishing among test takers. Kane (1996) noted,

The precision of change scores will depend on the intended interpretation of the change scores, in particular, on whether the interpretation focuses on the absolute magnitude of change for each individual or on the change for each individual compared with the average change in some reference group. (p. 368)

Thomas and Zumbo (2012) provided some support for the use of difference scores.

Raudenbush and Jean (2012) looked at reliability in the context of using value-added scores to examine teacher effectiveness, endorsing the use of confidence intervals over point estimates, and discussing multiple issues related to the use of value-added scores. Brennan et al. (2003) discussed various univariate and multivariate GT approaches to examine the reliability of group mean difference scores based on longitudinal data, from both norm-referenced and criterion-referenced perspectives, viewing the difference score as a composite score (the difference) of the two individual assessment scores.

Geldhof et al. (2014) examined reliability for data based on multistage sampling, concluding that level-specific calculations are appropriate when using multilevel data. Marcoulides (2019) examined reliability estimation in longitudinal studies, specifying the need to ensure that the approach used to model growth and measurement error aligned with the data being analyzed. Boyd et al. (2013) proposed an approach for estimating measurement error when students were administered three or more assessments, such as state assessments in consecutive grades. Their model allows for changes (increases or decreases) in knowledge and skills between administrations, as well as for the assessments to be neither parallel nor vertically scaled and to vary in their degree of measurement error.

Some research has examined change, difference, or growth scores in more complicated settings. Schuurman and Hamaker (2019) provided a preliminary model for some types of measurement error within the framework of autoregressive time series modeling, discussing how it could be relevant to both within-person and between-person precision. Tisak and Tisak (1996) used a latent curve approach to look at longitudinal models of reliability (see also Brandmaier et al., 2018).

DCMs have advanced to accommodate longitudinal data, examining how test takers' status changes over time. Madison (2019) defined and evaluated two reliability measures for longitudinal DCMs. The first measure is based on the idea of administering the same test repeatedly; the second is focused on the consistency of attribute mastery transitions, calculating the average most likely transition probability. Madison used both simulations and an empirical data set to study the two reliability metrics compared to a longitudinal IRT model. Consistent with the results of Templin and Bradshaw (2013),

the longitudinal DCMs had higher reliability than the longitudinal IRT models and were easier to interpret.

Brennan et al. (2003) focused on longitudinal difference scores for matched students. In practice, schools likely have three groups of test takers when looking at growth across grades: students present (and tested) in both grades, students present in only the lower grade, and students present in only the higher grade. For the test takers having assessment scores in both grades, their scores tend to be positively correlated, which tends to decrease the standard error associated with the mean difference scores, whereas this is not the case when the test takers in the lower and higher grade samples do not overlap. In addition to larger error, conceptually using independent samples to look at differences across grades makes drawing conclusions about instructional effectiveness questionable (Brennan et al., 2003), a reminder that the data collection design needs to be considered in terms of the questions one wishes to consider.

These more complex designs and analyses tend to introduce additional sources of error that need to be considered in addressing precision issues, some of which may not be estimable. Haertel (2013) discussed many issues to consider related to teacher value-added scores, and Kane (2017) laid out issues related to using residualized student gain scores for value-added models, including bias in the individual student residual gain scores and the impact of this bias when averaging across students to obtain estimates of a teacher- or school-level effect. A discussion of possible “corrections” is provided, but with the caution that estimates reflecting “all sources of random error in the prior scores” (Kane, 2017, p. 10) are typically not available.

Adaptive Testing

Computerized adaptive tests (CATs) differ from other assessments in that the items a test taker is administered depend on the test taker’s estimated proficiency. Multi-stage tests (MST) are typically of a fixed length, meaning a predetermined number of stages (and items) are administered to each test taker. Item or passage CATs generally stop after a set number of items are administered or after a fixed precision is reached. Although the standard errors of the proficiency estimates for each test taker can be used to obtain an approximate estimate of reliability, the standard errors will typically differ across test takers based on augmented stopping rules, such as stopping testing after a maximum test length is reached, regardless of the standard error of the current proficiency estimate.

A salient issue in adaptive testing is that basing which item or set of items a test taker sees next on how they responded to previous items suggests a violation of the item independence assumption, which IRT methods tend to rely on. The majority of methods used to estimate reliability for adaptive assessments use IRT methodology, though Divgi (1989) presented two methods to estimate reliability for a CAT without invoking IRT, and S. Kim and Livingston (2017) investigated a CTT-based procedure to estimate reliability for a MST, finding through their simulation that the method produced accurate results.

B. F. Green et al. (1984) stated, “Many psychometricians feel that devising a reliability coefficient for an adaptive test is inappropriate and misguided” (p. 352), but provided two alternatives, one being similar to an information function and the other using marginals. The authors suggested getting empirical reliability through test–retest designs. Beiser et al. (2016) computed test–retest reliability on an adaptive assessment of depression, concluding that the assessment provided “reliable screening.” Haley et al. (2010) reported a study where test–retest reliability was computed using intraclass correlations across four CATs.

The *Standards* (AERA et al., 2014) suggest estimating reliability and precision information through simulations. Thissen (2000, p. 166) concurs, stating that

because the entire system is used, including the item pool, the item selection algorithm, and the item exposure control system, such simulations may be expected to give accurate predictions about the performance of the CAT. . . . Simulation is the only situation in which the “real reliability” or “theoretical reliability” of a test can be determined.

Nicewander and Thomasson (1999) presented three test information–based reliability estimates for the Bayes modal estimate (i.e., MAP) and the ML estimate of proficiencies derived from direct definition, harmonic means, and Jensen’s inequality, demonstrating that the latter two estimates were upper bounds for the true reliability. The results of multiple simulations found that ML and MAP provided nearly identical true reliabilities in all data sets and that all reliability estimates were within .02 of the true reliabilities. Segall (2001) introduced two methods designed to improve measurement precision of “a general test factor,” one using a MIRT proficiency estimate and the other adaptively choosing items to maximize precision of a general proficiency score. Both methods were found to improve precision.

Nicewander (2018) transformed a reliability index to provide precision information for individuals, subgroups, and cut scores (conditional reliability coefficients), applying them to number-correct scores and theta estimates computed from number-correct scores, as well as theta estimates used as CAT scores. Seo and Jung (2018) compared three observed standard error marginalization methods (arithmetic mean, harmonic mean, and Jensen equality) in estimating empirical CAT reliabilities, finding that the Jensen equality method provided better accuracy and easy computation. Park et al. (2017) introduced a new method to obtain an analytic derivation of MST information.

Jodoin et al. (2006) studied test–retest and alternate forms reliability coefficients for several designs, including two- and three-stage MSTs, speculating that they would obtain higher coefficient values if there were more separations between their modules (i.e., if the medium module would have been less similar to the easy and difficult modules). Zhang et al. (2006) compared two IRT-based procedures for computing an aggregate reliability estimate using data from an MST assessment consisting of both MC and performance

tasks. One procedure used an empirical ability distribution to estimate score variance, whereas the other assumed a normal distribution. The authors concluded that the two methods tended to produce similar results. The authors raised the issue that if the two methods provide quite different estimates, the question to consider is “whether reliability estimation in IRT scored tests, especially adaptive tests, should be sample-driven or sample-free” (Zhang et al., 2006, p. 13).

One issue to keep in mind in MST is that test takers are routed to a subset of items, so computing precision information over all items in a panel falsely inflates precision because each test taker is administered only the items included in a single path, not all possible items included over all paths.

C. Wang (2014) combined adaptive testing with hierarchical latent trait estimation, in hopes of increasing the reliability of hierarchical latent trait estimation. Two item selection methods were proposed, and both improved measurement precision over a unidimensional item selection method, particularly for short tests and when the correlation between dimensions was high.

Schmitt et al. (2010) investigated how speededness affected the reliability of proficiency estimation in a CAT, using simulation and two item pools, one “real” and one “ideal.” The proficiency estimates became increasingly negatively biased as the CAT became more speeded in the real pool; these results were not replicated with the ideal pool.

Future precision research in adaptive testing will likely incorporate more complex IRT models, scoring with plausible values, and more complex assessment situations involving gaming and other simulations, where routing may be based on factors in addition to ability estimation.

Reliability Issues Related to Speeded Assessments

Operationally, some assessments are untimed, some are generously timed or partially speeded for at least some test takers, and still others have time limits imposed for practical reasons, such as when a test is administered in a school cafeteria and needs to be finished in time for lunch (for issues related to timing, see Margolis & Feinberg, 2020).

Estimating reliability for pure speed assessments from a single administration has been discredited for some time, with a consensus that administering two or more forms of an assessment to a group of test takers is the appropriate way to investigate precision (Anastasi & Drake, 1954; Cronbach & Warrington, 1951; Gulliksen, 1950), though attempts have been made to develop improved methods because of practical difficulties in obtaining data on two or more forms for a group of test takers (see Gulliksen, 1950). Cronbach and Warrington (1951) listed three conditions that result in inaccurate split-half or coefficient alpha reliabilities based on a single form when the assessment is speeded, including small variability in the number of items test takers have answered, when test takers have responded to all items they are likely to answer correctly within the time limit, and when the variation in the number of

items responded to from form/occasion to form/occasion is small compared to the number of items responded to across test takers. According to these authors, if data from an assessment with a time limit do not meet one or more of these conditions, a split-half reliability estimate from a single administration might be usable.

In achievement and licensure/certification testing, relatively few assessments are pure speed tests, meaning there is not an expectation that lack of time is the sole, or even primary, reason for test takers not answering every item correctly. According to Cronbach and Warrington (1951), speededness is a factor to consider when a test taker's position in a group of test takers would change if the test taker were given additional time to respond to the assessment. Speededness violates the assumption of local independence of items, in that for those items a test taker does not have time to answer, the lack of time affects the test taker not answering the item correctly more than other characteristics, such as item difficulty, thus impacting the assumptions both GT and IRT methods make about local item independence for estimating coefficients.

Attali (2005) used simulations to illustrate the impact of speededness on reliability, assuming that test takers who were running out of time would engage in rapid guessing on their remaining items as opposed to leaving them unanswered, which tends to lower reliability because some items are answered correctly by chance, which is unlikely to be consistent across multiple test forms. Whether the speed at which a test taker works is consistent across alternate forms is another consideration.

The salient point is to what degree the test taker differences in scores are due to speed, or what proportion of total score variance is "speed" variance (Attali, 2005; Anastasi & Drake, 1954). When assessments are slightly speeded, or only a small portion of total variance is due to speed factors, internal consistency coefficients may not be too misleading. However, with the increase in computerized testing and the collecting of timing information, such as the amount of time a test taker spends responding to each item, more sophisticated methods of addressing speededness are possible, such as explicitly including timing information in estimating proficiency. Petscher et al. (2015) used a conditional item response model that incorporates response time as an item parameter in the context of a reading fluency assessment. Their model subsumes the idea of independence at the individual level, but incorporates a joint relation between speed and accuracy at the population level. The authors found their model improved the precision of student scores by an average of 5%. The increase was not uniform, however, having the most improvement for higher proficiency students. Other models, such as those introduced by van der Linden (2009) that incorporate latency, could also be used.

Aggregation Issues

Reliability of Group Means

The reliability coefficients dealt with in this chapter have concentrated on individual test takers and precision related to normative or criterion-referenced issues related to individuals. There are, however, instances where the unit of interest is at the aggregate

level, for example, looking at a classroom. Examining trends, such as a cohorts' set of scores over time, or change scores, such as pretest and posttest scores to examine the effectiveness of a treatment, involves additional precision issues. Geldhof et al. (2014) examined level-specific (e.g., individual and group) reliability using three estimates and simulated and real multilevel data, finding support for the use of level-specific reliability estimates.

It is well known that the variability of, for example, individual fifth graders' assessment scores in a school district is generally greater than the variability of the average scores of the fifth-grade classrooms. Brennan (1995) used GT to illustrate, however, that the error variance for individuals could be less than the error variance for groups under certain conditions, depending on whether one considered persons and items random or fixed in terms of the universe of generalization. According to Brennan (1995), "Aggregation may well lead to a sizable decrease in error variance, but this can be very misleading if an investigator fails to take into account the corresponding decrease in true (or universe) score variance" (p. 395).

Reliability Generalization

Reliability is not a constant property of an assessment, but may vary based on context, such as a particular sample of test takers or number of raters, and the definition of a replication. Reliability generalization (RG) is basically a meta-analysis of reliability coefficient estimates across different studies, with the goal of estimating an average reliability across studies, as well as investigating the variability in the reported estimates and identifying factors that seem to be influencing the reliability estimates.

The importance of RG studies is that many researchers do not provide reliability estimates based on their particular context, but instead rely on, perhaps, a publisher's technical documentation for their reported reliability information, even though their context may differ in ways likely to affect precision. Although not as ideal as having precision information calculated on their particular data, RG analyses may provide some information on whether "borrowing" reliability information from one context and assuming it applies in another is likely to be reasonable (see Vacha-Haase et al., 2002, and Whittington, 1998).

There are different ways to implement an RG study and different factors to consider as possible influencers on estimated coefficient values. Holland (2015) examined 107 peer-reviewed RG studies, finding most analyzed coefficient alpha coefficients. He provided a set of guidelines for both conducting and reporting RG studies. Sánchez-Meca et al. (2019) also developed a checklist to assist researchers in reporting reliability generalization results. Henchy (2013) examined 64 RG studies, comparing the recommendations around such studies to how they were actually conducted, finding that some recommendations, such as conducting a priori power analyses, were not followed. Those conducting RG studies need to attend to issues such as whether to weight study results by test-taker sample size and how to treat a possible lack of

independence of reliability estimates where a given study has reliability estimates from various subgroups, subscales, or multiple methodologies.

Precision Issues Related to Raters

Rater Reliability

When the scoring of a test taker's responses is impacted by who does the scoring, this source of variability should be incorporated when computing reliability. Rater impact is often a matter of degree and encompasses various situations, such as a teacher and a teacher's aide interpreting penmanship differently to various trained raters scoring constructed response items using an elaborate rubric on a statewide accountability assessment. The more likely different raters are to assign the same score to a test taker's response, and the more likely the same rater is to assign the same score to the same response, the less rater variability contributes to error variance. Good rater training, well-defined scoring rubrics, and monitoring to ensure raters are interpreting the rubrics and responses the same way help increase rater consistency.

GT is frequently used to model the variability of raters, as well as to estimate precision information over various features, such as the number of raters and the number of prompts or tasks administered. However, it is not always clear how well the results of these studies generalize to a particular context. Studies done by publishers during the initial administration of an assessment may include expert raters, training materials that are still under development, and prompts and rubrics that are still being tweaked, as well as administration conditions, such as time limits, that are still being determined, and they may or may not include test takers who are motivated. All of these factors can impact the GT precision information and make it less relevant for a later context. In addition, under operational conditions, raters are likely not all at the expert level, and raters may not be assigned randomly but rather by convenience, such as which raters happen to be available for a particular shift when a particular set of responses are available for grading. For local GT analyses, existing data are often searched for a small subset of data that align with a standard GT design, such as a fully crossed rater by test taker nested within prompt design, to gather precision information, including rater effects.

Rater scores are often compared through either a rank ordering of responses (correlation between rater scores over a group of test-taker responses) or a match of actual assigned scores (often contingency tables focusing on exact or adjacent matched scores, frequently corrected for chance agreement). Rater severity and leniency; the tendency of a rater to assign high or low scores, respectively; and proclivity of a rater to use the whole range of scores or not are perhaps the most frequently cited rater effects. The scoring process logistics, such as scoring all responses to a single prompt versus scoring all responses by a single test taker, can mitigate some rater issues, such as the halo effect.

Studies should account for variability within and across raters. Models incorporating rater effects include IRT models (Linacre, 1989; Lunz et al., 1990; Verhelst & Verstralen,

2001), a hierarchical rater model (Patz et al., 2002), a signal detection theory model (DeCarlo, 2010; DeCarlo et al., 2011), Yao's rater model (Z. Wang & Yao, 2013), and Longford's (1994) model-based approach for multiple raters of essays. Recent research includes that of Zapf et al. (2016), who investigated confidence intervals for nominal data and Fleiss's kappa and Krippendorff's α for interrater reliability in multiple scenarios, finding that bootstrapped confidence intervals provided coverage probability close to the theoretical ones. Choi and Wilson (2018) proposed a generalized linear latent and mixed model approach, combining IRT and GT, which allows ML estimation of individual random effects and variance components for generalizability coefficients. Gianinazzi et al. (2015) asked the same raters to rate the same responses at two points in time, looking at both intrarater and interrater reliability. Abdalla (2019) looked at the scoring of prompts reused over time under different conditions of rater drift, looking at the percentage of exact agreement and Cohen's kappa as interrater agreement measures, and the paired t test and Stuart's Q as marginal homogeneity measures, under two data frameworks: the generalized partial credit model and latent-class signal detection theory model.

Haertel (2006) discussed several rater issues, including what is the true score of test takers on a prompt where raters score responses. Is it the average score across replications of raters? Or across a test taker's repeated responses to the prompt, if the test taker could be repeatedly administered the same prompt? Or is it based on the rubric and not on rater's assigned score? Haertel (2006) also mentioned issues related to scoring with a finite rubric, such as 0 to 3, and the implications for test takers with true scores at the extremes. For example, a test taker with a true score of zero can only have measurement errors in one direction. Haertel (2006) stated, "Unless all raters assign the identical score, the mean observed score over raters is a biased estimator of the essay's true score. Thus, from this perspective, true score cannot be equal to the expected value of observed score" (p. 102). Haertel emphasized the need for precisely specifying the definition of true score and error, as well as the model being proposed, to minimize ambiguity and determine the appropriateness for a particular context.

Haertel (2006) also raised the issue of what score is assigned when multiple raters score the same response. Typically, multiple scores are averaged only if the scores are identical or fall within some predetermined range of each other; for example, adjacent scores may be averaged, but scores that are not identical or adjacent may be sent to resolution. Resolution may involve ignoring the previous rater scores and using only a score assigned by a more expert rater. Other rules may involve keeping the original rater score that most aligns with an expert score, discarding the more disparate original rater score. It is also the case that any averaging of ratings is likely to result in fractional scores that do not directly align with the scoring rubric. Both fractional scores and resolution scores impact variability due to raters. Haertel (2006) suggested that additional research on how to incorporate these cases into a rater model is warranted.

Variability across raters can be examined by assigning multiple raters to score the same responses. Variability within a rater is somewhat more difficult. A rater may

remember seeing a particular response previously, so having a rater score a single response multiple times as independent scores may be challenging, as well as resource intensive. Examining how consistently a rater scores “essentially identical” responses might provide some insight, though the issue then becomes what is essentially identical and under what conditions (time of day, position in a stack of responses, before or after a higher or lower scoring response, length of time between scorings, etc.). Though raters are typically not the largest source of error in the precision of rater-scored responses, lagging behind task/prompt variability, rater variability is still a source of unreliability to minimize.

Automated Scoring Issues

Machine scoring, or artificial intelligence (AI) scoring, involves training a computer, through natural language processing or other methodology, to score essay/constructed responses by machine, rather than human raters. AI scoring is appealing because, in theory at least, it is more efficient and returns scores faster and more cost-effectively than human raters. Note, however, that current AI engines tend to require some training on each individual prompt, so the benefit in efficiency is largely manifest at scale when large numbers of test takers respond to the same prompt. (See, however, Attali, 2011 and Foltz et al., 2013 for a generic model; Foltz et al., 2013 reported the generic AI rater as having about 10% lower reliability than the prompt-specific rater). Although most of the research literature in AI seems to focus on essay scoring, there have been studies on other types of constructed response items (see Shermis, 2015, who looked at machine scoring short-answer items).

Whether assessment responses are scored solely by a single AI engine, multiple AI engines, or a combination of an AI engine and one or more human raters, precision coefficients are impacted. The most common way AI scores are currently evaluated is to look at the consistency between scores assigned by human raters to a group of test-taker responses correlated with the scores an AI engine assigns, where the AI engine scoring is typically trained to maximally predict the human scores based on features that may or may not resemble the human scoring rubrics. Various coefficients, including Pearson correlations between the two sets of scores, exact and adjacent agreement, and kappa, have been used (see, for example, Attali, 2013, who examined correlations between AI and human scoring across multiple studies, typically comparing human–human correlations to AI–human correlations). Attali (2015) examined different weights applied to AI features to find the optimal weights for reliability and validity, as opposed to the traditional use of predicting scores human raters would assign, reporting that the feature weights that predicted human rater scores were different from those that emphasized reliability coefficients.

Attali and Burstein (2006) discussed issues related to using human rater and AI agreement as reliability, including when data on only a single prompt are used. While a human rater may or may not assign the same score to a test-taker response presented twice, an AI engine would be expected to return the same score, assuming that no

intervening tweaking or training of the engine occurred, suggesting it may be preferable to compare scores across multiple prompts rather than a single prompt.

Attali et al. (2013) looked at increasing the reliability of combined human and AI scores, both by decreasing the overlap in the features rated and by increasing the reliability of the human scores. Both methods proved effective. Attali (2011) reported that a test–retest reliability for a single human rater was .53, while that of a generic AI rater was .80. However, in the same way that all prompts, rubrics, administration conditions, samples, human rater training protocols, and so on vary, so do different AI engines, so while these results may be informative, they may not generalize to other contexts.

One question that arises in considering AI scoring is what actually constitutes a replication. Sending a specific response through the same AI scoring engine twice does not reflect the same variability as sending the same response through human scoring twice, with either the same or a different rater. The same test taker could be administered the same prompt or form twice and both responses could be sent through the AI engine, but this is not really a replication to investigate the AI engine because the within-person variability confounds the interpretation. Perhaps training the AI engine twice on two separate sets of training or calibration papers and then running a new set of responses through both engines and comparing the paired scores would be informative, but still would not really align with the typical human rater analyses.

Reliability of Behavioral Observations

Another setting where raters are involved is in observations of behavior, such as observers in a classroom setting looking for incidences of certain behaviors, the absence or presence of which on the part of teachers and/or students is recorded. Individual rater errors consist of recording an occurrence when it is not manifest and not recording an occurrence when one is manifest. Issues in this context involve rater consistency (one rater observes a behavior and one does not) and rater accuracy (either the behavior did or did not occur, so both raters observing the same incident are not correct), and representativeness—if one wishes to make inferences about behaviors based on the observed/rated sample, how representative is the observation period? For example, is an observed teacher demonstrating typical behavior, or is the fact of being observed leading a teacher to be more likely to demonstrate a desired behavior?

Gwet (2012) discussed issues around interrater reliability in medical settings, raising the issue of whether two raters whose ratings correlate highly are “interchangeable” and to what extent we can extrapolate from raters in a study rating particular subjects to a broader context, emphasizing the importance of gathering precision information using raters similar to those that will be used operationally. Gwet (2014) also discussed ordinal measurement issues, when the distinction of whether raters agree or disagree may be a matter of degree as opposed to absolute, offering the illustration,

Two raters A and B who rate the same patient as “Certain Multiple Sclerosis” and “Probable Multiple Sclerosis” are not quite in total agreement. But are they in disagreement? Maybe to some extent only. That is, with ordinal scales, a disagreement is sometimes seen as a different degree of agreement. (p. 16)

This suggests that a rank ordering of test takers, rather than a contingency table, might be preferred to examine consistency.

The reader is referred to Haertel (2006) for additional information on issues related to precision with behavioral observations, particularly his treatment of Rogosa and Ghandour (1991) and their framework for considerations of score accuracy.

Error in Reliability Estimates

Reliability coefficients are used to provide an indication of the uncertainty in measurements (see Frank, 2002, for the use of reliability coefficients to correct effect sizes), but reliability estimates also contain error. In the examples provided in W. Lee et al. (2025), coefficient values across the two contrived forms can be used as a measure of consistency, but in practice, it is uncommon for more than one administration to occur, and simulations or resampling procedures are often used to estimate standard errors and confidence intervals. While this information can be informative, some sources of error, such as test-taker consistency across alternate forms or separate administrations, are not fully considered.

Several researchers have investigated issues around error in reliability and precision estimates. Andersson and Xin (2018) derived asymptotic variances for IRT marginal and test reliability coefficient estimates for both dichotomous and polytomous IRT models and demonstrated through simulations that the resulting confidence intervals had “good coverage” under several conditions studied, with the marginal reliability coefficient having somewhat lower sampling variability and larger bias than the test reliability coefficient.

Feldt (1990) provided methodology to create confidence intervals for the intra-class reliability coefficient, used when there is a single rater for individual responses, based on sampling theory. Ogasawara (2009) derived asymptotic distributions of several sample coefficients with and without stratification under nonnormality. Terry and Kelly (2012) focused on estimating the sample size needed to obtain narrow confidence intervals for composite coefficients. Bujang et al. (2018) looked at sample sizes relative to hypothesis tests for Cronbach’s alpha coefficient. Zapf et al. (2016) looked at confidence intervals for interrater reliability estimates for nominal data.

When using an estimate of precision, particularly to inform a high-stakes decision, it is important to consider the precision of the estimate as well. This is likely to be of particular concern in GT where variance component estimates may be based on a sampling of very few levels from an infinite population (see Brennan, 2001c).

Practical Guidelines

What Reliability Computations Are Most Appropriate for What Context?

As has been discussed throughout this chapter, there are a variety of methodologies that can be used to compute precision information. According to the *Standards* (AERA et al., 2014), the reliability methodology selected for use in a particular context should “be consistent with the structure” of the assessment, providing the example that an assessment with considerable multidimensionality would indicate that a composite score reliability be used. It is also recommended that precision information be computed for the scores on which decisions are going to be made. For example, if equated and rounded scale scores are used for placing students into their first-year college math class, precision information provided on equated and rounded scale scores will be more appropriate than internal consistency information on raw scores.

E. Haertel (personal communication, March 8, 2019) advocated framing the issue of which reliability computations are most important by first considering what the intended use of the data is and what sources of error should be considered (“begin with a substantive definition of the intended universe of generalization”). This is then followed with selecting a data collection design that would populate the appropriate model. GT can often be extremely helpful here by explicitly identifying what factors one wishes to generalize over calculations (see Revelle & Condon, 2019). It is not always possible to collect the desired data, for example, administering the same assessment over and over to the same test takers. However, knowing what is desired can be helpful in choosing what available methodology to substitute.

To explore the various reliability statistics that can be used to represent consistency and inconsistency of scores, two large-scale examples are provided by W. Lee et al. (2025). For each of the examples, multiple reliability coefficients, CSEMs, and classification consistency and accuracy coefficients are provided. In addition, results are provided for number-correct raw score, IRT proficiency, and scale-score metrics for various statistics and indices. The purpose of providing these examples is to illustrate similarities and differences in values across multiple reliability statistics and to reinforce the importance of identifying details about how a reliability estimate was computed, instead of simply reporting “the reliability was [some value].”

W. Lee et al. (2025) demonstrated through examples that different reliability coefficients calculated on a single set of data yield different values. This is largely because the different coefficients make different assumptions. Simply running a multitude of analyses and choosing to report the highest values is counterproductive. Woodruff and Wu (2012), in a paper comparing coefficient alpha with multiple other coefficients, reminded us to remember what each reliability coefficient indicates and that we should be reporting the coefficient that best aligns with the inferences we plan to make.

How Do I Increase Reliability and How Much Do I Need?

The *Standards* (AERA et al., 2014) do not provide criteria for minimally acceptable reliability coefficients. Occasionally a request for proposal, a government organization, or an author (see Dimitrov et al., 1999) may specify a particular value of reliability that should or must be achieved, but there is rarely any rationale provided for the chosen value or any criteria for the conditions under which the value must be achieved (such as test-taker sample size, whether under operational or special study conditions, or the metric or type of coefficient). Reliability coefficients appearing in technical manuals may provide some indication of the magnitude thought acceptable, but typically these are lacking in detail about the data collection, and there is no evidence that these values are “adequate” other than that the test publisher thought they were sufficient to publish the assessment. Ellis (2013) commented, “The absence of such standards is a serious gap in existing behavioral research methods and leads one to ask why reliabilities are routinely computed if their acceptable values are unknown” (p. 16). Wainer and Thissen (1996) brought up the related issue of the practical impact of small differences in reliability coefficients.

Kane (1996) advocated judging how adequate a precision level is for one’s intended use by considering the measurement error in relation to the “tolerance for error” in a specific situation:

The level of error that is tolerable in a particular context is determined by the interpretations to be applied to the measurements and the uses to be made of the measurements. To the extent that errors do not interfere with intended interpretations or uses, they are not a serious problem. (p. 356)

This leads to reframing the question “How much reliability do I need?” to “How much error (uncertainty) can I tolerate in the decisions I need to make with these data?” Practitioners may not be concerned with relatively large CSEMs occurring far away from a cut score. However, when using an assessment as a component in a decision, such as a test score, a grade point average, or soft skill measures in determining college admissions, it may be much more difficult to articulate the amount of “tolerance for error” in a component.

Different factors impact different precision calculations differently, again depending on what the specific calculations account for. For example, differences in test specifications across test forms have no impact on computing a reliability coefficient based on only one administration of a single form, but would have an impact on a coefficient involving parallel forms or on analyses involving different forms over time. Alternate forms reliabilities will tend to increase as the forms are more similar.

Internal consistency measures increase as the items within an assessment have higher discrimination values. To rank order test takers according to some characteristic or trait, items that all test takers or no test takers answer correctly provide no information. Jodoin et al. (2006) reported that higher coefficient values for MST tests would be obtained if there was more separation between the easy, medium, and difficult modules.

The more objectively an assessment or item is scored, the more likely the reported scores are to be consistent and the less the scoring process is likely to contribute to error variance. The way an assessment is scored can affect reliability; Pugh and Brunza (1975) stated that the reliability of a test increased from .57 to .85 when comparing traditional number-correct scoring to a confidence scoring method. In addition, when multiple components are used to form a composite score, giving more weight to the more reliable components tends to increase reliability of the composite. Kane and Case (2004), for example, addressed how to weight two components of a composite score to maximize correlation with desired true score (defined as a specified weighted composite of component true scores) when the observed component scores have different reliabilities.

Another factor that can impact reliability coefficients is the metric used, as shown in tables associated with the two examples reported by W. Lee et al. (2025). There are multiple ways in which scale scores, theta scores, and cut scores could be set, which would also potentially influence precision calculations.

One factor that seems to be generally accepted is that longer assessments have more precision than shorter assessments because they provide a larger sample of tasks or items. However, there are caveats to keep in mind. One is that increases in precision due to increasing assessment length is likely most effective when the initial test is short. Adding 30 items to a 10-item test typically will increase reliability more than adding 30 items to a 200-item test, other things being equal. “Other things being equal” is an important caveat. Adding poor items (where “poor” may indicate inferior statistically, or in terms of alignment to a test blueprint, or items of such novel format that test takers are confused, or items that make the test so long fatigue and lack of motivation become issues) are more likely to lower precision than increase it. As Thorndike (1951) stated, “It has been shown that the reliability of many tests could actually be increased by omitting a number of items on the test” (p. 602).

In addition to increasing the number of items, the length of a multiple-choice assessment may also be considered in terms of increasing the number of options per item, again assuming the additional options are viable and of equivalent quality. The relationship of item sets to reliability may be somewhat more complex because of local dependence. Adding additional items per item set, where all items refer to the same set of stimulus material, is unlikely to have the same impact on reliability as adding additional item sets. According to Livingston (2018), “A test taker who has difficulties with that particular stimulus will have trouble with all the questions about it. To improve alternate-forms reliability, we need to increase the number of item sets” (p. 16).

Wainer and Thissen (1996) provided an example for which a test of shorter passage sets needed 43% additional items and a test of longer passages need 65% more items for each to obtain the reliability of a test of 40 stand-alone items. Lawrence (1995) also looked at the issues of local item dependence and the effect on reliability, finding that item sets effectively reduced test length: “The effect of item sets is revealed by a systematic discrepancy between the internal consistency estimates, which treats items based

on the same passage as independent, and the parallel-form estimates” (p. 13). Passage sets may be an excellent way to assess certain content, but the introduction of item dependence affects some precision coefficients through a violation of assumptions.

When scores involve raters, adding additional independent raters tends to increase precision, again assuming the additional raters are of similar quality as the original raters. The question of whether adding additional prompts/items or additional raters would be most likely to improve reliability in a particular setting requires the total design to be considered. GT studies allow one to estimate the influence different factors would have on certain types of reliability coefficients, such as whether adding additional items or prompts, raters, clusters of students per school or more schools, and so on would provide a better return on investment in a particular context (see Sun et al., 1997). The design implemented for data collection impacts precision, especially from the standpoint of the intended universe of generalization. Whether a particular set of raters or prompts is considered fixed or random may have a substantial impact on the coefficient values. However, the goal is not to choose a design, such as assuming that raters are fixed, to raise a coefficient value, but to have the design reflect the context one is interested in.

Different data models, as well as different precision estimates, have different assumptions and incorporate different sources of variability into their definition of error. Therefore, selecting a coefficient because it is larger is not recommend. Coefficients should be selected to fit one’s desired use and to align with the data collection design one implements.

Inappropriate administration conditions for an assessment can lower precision estimates by adding in additional error. These factors can range from inappropriate time limits, use of unfamiliar item formats or technology, inadequate motivation on the part of test takers to score their best, inadequate test directions, sloppy scoring, and a myriad of other conditions. Adequate instructions and practice exercises to ensure test takers understand what they are being asked to do tends to increase precision (see Thorndike, 1951).

For some consistency measures, the test-taker sample on which they are computed is a factor with large influence on the calculated value of the coefficient, such as the distribution of masters/nonmasters for a classification consistency measure or the overall distribution of test takers, such as how bunched or spread out the test takers are in terms of the underlying trait one is trying to estimate. The issue of restriction of range, when the possible value range is truncated through self-selection (for example, when only previously failing candidates are administered a retest or only high-proficiency candidates self-select to take an assessment) has been studied by various researchers. Sackett et al. (2002) examined the effect of different scenarios of range restriction on reliability coefficients, finding that the different scenarios resulted in different underestimates of criterion reliability. Fife et al. (2012) also looked at various selection ratios in relation to several reliability coefficients, including KR20 and test–retest, corrected and uncorrected, under varying selection ratios. These estimators were examined in

terms of bias and precision, with test–retest reliability usually being the best estimator of reliability across conditions.

Improving precision has multiple perspectives. Deliberately testing a sample with an extreme range of ability may result in a larger calculated value of a coefficient for, say, technical documentation, but doing so influences only that particular administration, whereas adding additional items to the test will likely improve precision for every future administration. Similarly, narrowing the desired universe of generalization, from a GT perspective, by considering, for example, raters fixed may raise a value, but is inappropriate if those raters will not be scoring future responses. Knowing what factors may influence a particular reliability coefficient facilitates comparing values across assessments by considering the context under which each coefficient was calculated.

Effective Ways to Communicate Reliability Information

Once reliability information has been estimated, these values need to be communicated to those wanting to make decisions based on the scores. The statement “The reliability of Assessment A is .80” is virtually meaningless in terms of helping a user know how much confidence to place in the scores of interest. Without additional information, one cannot even be sure that a value of “.80” is better than a value of “.70” but not as good as a value of “.90,” because one does not know whether the three values are for the same type of coefficient (or what sources of error were incorporated), or what sample of test takers was used in each instance, or other details, such as administration conditions, scoring, and so on. When presenting precision information, the relevant context must also be presented, so in addition to information about the actual coefficients computed, information on the assessment, the administration, the test-taker sample, the scoring, the reported score metric, and other pertinent details should be included. An internal consistency coefficient of a released practice form given under lax conditions with no stakes to students reported on number-correct scores and an alternate-forms coefficient under high-stakes standardized conditions to a self-selected sample reported on a scale score metric might be interpreted quite differently by someone planning to incorporate the assessment into a composite for decision-making, even if the coefficient values were similar.

Given there is sufficient detail to understand the context of the design and the coefficient calculations used, “uncertainty” needs to be communicated in a manner that facilitates understanding. A statement such as “A coefficient value of .75 suggests that 75% of the variance in scores is due to actual differences in ability while the remaining 25% is attributable to measurement error” is unlikely to be helpful to a lay audience. S. Johnson and Johnson (2009, p. 51) stated,

If the public is to be educated about technical issues in assessment, and if reliability information is to be routinely published alongside examination results, then we need to decide which form of reliability measure would be the most appropriate one to use. There are basically two choices: a variance ratio (reliability coefficient), and a standard error of measurement.

According to C. Wang (2014, p. 454), “reliability has the advantage of being a compact measure of precision that has a fixed metric (bounded between 0 and 1) that is widely understood.” Cronbach (2004) asserted that the SEM should take precedence as the most important piece of information to report, rather than a reliability coefficient. Patterson (1955), in an article dedicated to interpreting SEMs, discussed problems in making statements about SEM concerning true scores using the standard error of observed scores.

Walsh et al. (2014) illustrated reporting information graphically rather than as a single number. Raudenbush and Jean (2012) advocated the use of confidence intervals for value-added scores for teachers; they also discussed issues around interpretation, such as shorter intervals (e.g., 75% versus 95% confidence intervals) and the different interpretations one might make (such as “Where do I as a teacher stand among all teachers?” versus an administrator wanting to identify a lower group of teachers for training or a higher group for recognition).

Regardless of one’s view on what measure of precision to report, illustrations and examples can help practitioners understand abstract concepts around uncertainty. For example, statements suggesting how many masters/nonmasters are likely to be misclassified, or how a test taker is likely to score if given the assessment a second time, or illustrating through simulation the distribution of obtained scores a test taker (simulee) of a given true proficiency achieves on two different CAT designs may be more impactful than simply reporting a coefficient value. Providing a comparison can also be useful to aid in interpretation, as Wood (2020) illustrated:

When possible, documentation should include agreement statistics for two independent human raters and for a human rater with automated scoring. The public may be surprised if a scoring engine agrees with a human 65% of the time but may not be so surprised if two independent humans agree only 67% of the time on the same prompt. (p. 144)

FUTURE DIRECTIONS

In the years since the fourth edition of *Educational Measurement* was released, there has been progress in addressing reliability issues with more complex data models, with incorporating additional sources of error in coefficients, and with the development of new methodologies and coefficients. There has also been increased emphasis on the use of CSEMs over reliability coefficients, which this chapter hopefully reinforces.

As computing power and the integration of computer science and data analytics into measurement approaches continue to advance, more complicated simulation methods are available to examine reliability across more complex assessment models. As our models and assessments become more complex, so do our methods of estimating reliability. From fixed forms to adaptive tests to game-based assessment experiences, the concept of replication becomes more challenging. It may be relatively easy to imagine a test taker taking the same fixed form (or a randomly parallel or tau-equivalent form) a second time.

However, when it comes to adaptive testing and game-based assessment, with various possible paths involving the same or different pools, the scenario becomes more complex.

Simulations have become more necessary to estimate reliability, such as running the same simulees through alternate CAT pools multiple times. These simulations rest in part on having realistic models to generate test-taker responses, and we are not there yet. Too many simulations are conducted with models with too small of a random (or even some systematic) error component. We need new models of data generation along the lines of those used by Kolen and Harris (1987) and Davey et al. (1997) to more accurately reflect actual test-taker behavior if we are going to rely on the generated data for reliability information.

The previous editions of *Educational Measurement* have primarily focused on the reliability of summed scores from a single point in time (Haertel, 2006). This chapter extends its focus to include additional metrics such as IRT proficiency estimates and scale scores, recognizing the importance of reliability information aligned with the specific metric used for decision-making. In the future, additional metrics will likely become more prevalent in operational use, requiring further advancements in reliability estimation. Although the adopted definition of reliability in this chapter is replication based, the focus remains on discrete points in time rather than the continuous testing model mentioned by Haertel (2006). These types of assessment models, familiar to some as ongoing formative assessments, still require additional development in terms of reliability information across all previously collected data, the definition of what a replication consists of in a model continually adding additional assessment information, as well as what we are attempting to make inferences about (and when) in an ongoing data stream.

Additional considerations warrant attention, including the exploration of whether automated item generation and item cloning can accurately predict item parameters or characteristics to a sufficient extent for computing CSEMs prior to administering an assessment. Moreover, evaluating reliability around test takers using plausible values, imputed item scores, automated scoring, and scoring that incorporates additional information from item latencies or item clicks is essential. It is particularly important to define the relevant error components and determine the composition of a replication in these scenarios.

Additional ways to communicate what different levels of reliability and CSEM translate to in practice are needed. For example, under classification consistency/accuracy, informing users of the likely percentage of test takers being misclassified either at a cut score or overall helps a user interpret what a difference in classification consistency/accuracy might mean in practice. Similar ways of translating or illustrating the impact of different levels of reliability for a particular inference would be beneficial. Because settings differ widely in the type of decisions one is making, for example, whether to have a student do additional work with finding common denominators prior to moving on to subtraction involving mixed fractions versus licensing an applicant to practice in medicine or law, having one-size-fits-all requirements on the level of reliability needed is inappropriate. However, providing guidelines on presenting reliability information to

enhance users' understanding of what a particular level of reliability means in a specific context would be helpful. In addition, more guidance on which types of coefficients on which metrics ought to be reported in different contexts would seem appropriate.

Cronbach stated in 2004,

I am convinced that the standard error of measurement . . . is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report on the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported. (p. 413)

Cronbach's plea warrants serious attention.

Since the fourth edition of *Educational Measurement*, there has been relatively little literature published on the integration of IRT, GT, and CTT, despite Haertel (2006) and others encouraging it. Hopefully, this chapter is a meaningful response to that goal, though there is still plenty of room for additional progress. Also, while there has been some progress on integrating the concepts of reliability and validity, to echo Haertel (2006), much remains to be done.

ACKNOWLEDGMENTS

The authors express gratitude to Edward H. Haertel at Stanford University, Seonghoon Kim at Hanyang University, and Feifei Ye at RAND for their invaluable contributions as reviewers, providing excellent and thorough reviews. They also extend thanks to Robert L. Brennan at the University of Iowa and Stella Y. Kim at the University of North Carolina at Charlotte for their valuable comments on the initial draft. Special appreciation is given to Huan Liu and Kuo-Feng Chang, graduate students at the University of Iowa, for their assistance with the literature review and data analysis.

REFERENCES

- Abdalla, W. (2019). *Detecting rater effects in trend scoring* [Doctoral dissertation, University of Iowa]. <https://doi.org/10.17077/etd.qh0s-2ij2>
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Anastasi, A., & Drake, J. D. (1954). An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement*, 14, 529–540.
- Andersson, B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational and Psychological Measurement*, 78, 32–45.
- Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, 18, 1–14.

- Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement*, 29, 357–368.
- Attali, Y. (2011). *Automated subscores for TOEFL iBT® independent essays* (ETS Research Report No. 11-39). ETS. <http://dx.doi.org/10.1002/j.2333-8504.2011.tb02275.x>
- Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.
- Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied Psychological Measurement*, 39, 303–313.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 1–29.
- Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30, 125–141.
- Bacon, D. R., Sauer, P. L., & Young, M. (1995) Composite reliability in structural equations modeling. *Educational and Psychological Measurement*, 55, 394–406.
- Beiser, D., Vu, M., & Gibbons, R. (2016). Test–retest reliability of a computerized adaptive depression screener. *Psychiatric Services*, 67, 1039–1041.
- Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement*, 17, 323–349.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.
- Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, 26, 364–375.
- Boyd, B., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics*, 38, 629–663.
- Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis. *Frontiers in Psychology*, 9, 294.
- Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, 32, 385–396.
- Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, 22, 307–331.
- Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38, 295–317.
- Brennan, R. L. (2001b). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, 20, 6–18.
- Brennan, R. L. (2001c). *Generalizability theory*. Springer-Verlag.
- Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). American Council on Education/Praeger.

- Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24, 1–21.
- Brennan, R. L. (2022). Generalizability theory. In B. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 206–231). Routledge.
- Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277–289.
- Brennan, R. L., & Kane, M. T. (1977b). Signal/noise ratios for domain-referenced tests. *Psychometrika*, 42, 609–625.
- Brennan, R. L., Kim, S. Y., & Lee, W. (2022). Extended multivariate generalizability theory with complex design structures. *Educational and Psychological Measurement*, 82, 617–642.
- Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement*, 59, 5–24.
- Brennan, R. L., & Lee, W. (2006). *Some perspectives on KR-21* (CASMA Technical Note No. 2). Iowa City: University of Iowa.
- Brennan, R. L., & Lee, W. (2008). Correcting for bias in single-administration decision consistency indices. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 15–24). Universal Academy Press.
- Brennan, R. L., Yin, P., & Kane, M. T. (2003). Methodology for examining the reliability of group mean difference scores. *Journal of Educational Measurement*, 40, 207–230.
- Breyer, F. J., & Lewis, C. (1994). *Pass–fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). ETS.
- Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296–322.
- Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: A simple guide for researchers. *The Malaysian Journal of Medical Sciences*, 25, 85–99. <https://doi.org/10.21315/mjms2018.25.6.9>
- Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245–314). The Macmillan Company.
- Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology*, 8, 103–118.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16, 221–248.
- Cheng, Y., Yuan, K. H., & Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educational and Psychological Measurement*, 72(1), 52–67.
- Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling*, 60, 53–80.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

- Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64, 391–418.
- Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, 16, 167–188.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97–4). ACT.
- DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR-10-08). ETS.
- DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48, 333–356.
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43, 145–168.
- DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13, 354–378.
- Dimitrov, D., Rumrill, P., Fitzgerald, S., & Hennessey, M. (1999). Reliability in rehabilitation measurement. *Work*, 16, 159–164.
- Divgi, D. R. (1989). Estimating reliabilities of computerized adaptive tests. *Applied Psychological Measurement*, 13, 145–149.
- Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika*, 16, 407–424.
- Ellis, J. L. (2013). A standard for test reliability in group research. *Behavior Research Methods*, 45, 16–24. <https://doi.org/10.3758/s13428-012-0223-z>.
- Feldt, L. S. (1975). Estimation of the reliability of a test divided into two parts of unequal length. *Psychometrika*, 40, 557–561.
- Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, 44, 883–891.
- Feldt, L. S. (1990). The sampling theory for the intraclass reliability coefficient. *Applied Measurement in Education*, 3, 361–367.
- Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). American Council on Education and Macmillan.
- Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement*, 33, 141–156.
- Feldt, L. S., & Qualls, A. L. (1998). Approximating scale-score standard error of measurement from the raw score standard error. *Applied Measurement in Education*, 11, 159–177.

- Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, 9, 351–361.
- Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, 31, 525–543.
- Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of α , ω , and test–retest reliability for dichotomous data. *Educational and Psychological Measurement*, 72, 862–888.
- Foltz, P. W., Streiner, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68–88). Routledge.
- Frank, B. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, 62, 254–263.
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19, 72–91.
- Gianinazzi, M. E., Rueegg, C. S., Zimmerman, K., Kuehni, C. E., & Michel, G. (2015). Intra-rater and inter-rater reliability of a medical record abstraction study on transition of care after childhood cancer. *PLoS One*, 10(5), 1–13. <https://doi.org/10.1371/journal.pone.0124290>
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423–436.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31, 4–19.
- Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika*, 48, 99–111.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–521.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement*, 66, 930–944.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347–360.
- Green, S. B., & Yang, Y. (2008). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, 74, 155–167.
- Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika*, 15, 259–269.
- Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment Research & Evaluation*, 11(6).

- Guttman, L. A. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, 10, 255–282.
- Gwet, K. L. (2012). *Handbook of inter-rater reliability* (3rd ed.). Advanced Analytics.
- Gwet, K. L. (2014). *The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). American Council on Education/Praeger.
- Haertel, E. H. (2013, March). *Reliability and validity of inferences about teachers based on student test scores*. The 14th William H. Angoff Memorial Lecture was presented at The National Press Club, Washington, DC. <https://www.ets.org/Media/Research/pdf/PICANG14.pdf>
- Haley, S. M., Chafetz, R. S., & Tian, F. (2010). Validity and reliability of physical functioning computer adaptive tests for children with cerebral palsy. *Journal of Pediatric Orthopaedics*, 30, 71–75.
- Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80–123). The Johns Hopkins University Press.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). American Council on Education and Macmillan.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 10, 159–170.
- Handcock, G. R., & An, J. (2018). Digital ITEMS module 2: Scale reliability in structural equation modeling. *Educational Measurement: Issues and Practice*, 37, 73–74.
- Hanson, B. A. (1994). *An extension of the Lord–Wingersky algorithm to polytomous items* [Unpublished research note]. ACT.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345–359.
- Henchy, A. M. (2013). *Review and evaluation of reliability generalization research*. [Doctoral dissertation, University of Kentucky]. https://uknowledge.uky.edu/edp_etds/5
- Holland, D. F. (2015). *Reliability generalization: A systematic review and evaluation of meta-analytic methodology and reporting practice* [Doctoral dissertation, University of North Texas]. <https://digital.library.unt.edu/ark:/67531/metadc822810>
- Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6, 153–160.
- Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 13, 253–264.
- Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics*, 15, 353–368.

- Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement*, 10, 175–186.
- Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement*, 6, 161–171.
- Jarjoura, D., & Brennan, R. L. (1983). Multivariate generalizability models for tests developed according to a table of specifications. In L. J. Fyans (Ed.), *New directions for testing and measurement: Generalizability theory: Inferences and practical applications* (No. 18, pp. 83–101). Jossey-Bass.
- Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19, 203–220.
- Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, 55, 635–664.
- Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Office of Qualifications and Examinations Regulation.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- Jöreskog, K. G., & Sörbom, D. (2018). *LISREL 10 for Windows* [Computer software]. Scientific Software International.
- Kane, M. T. (1996). The precision of measurement. *Applied Measurement in Education*, 9, 355–379.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education/Praeger.
- Kane, M. T. (2017). *Measurement error and bias in value-added models* (ETS Research Report No. RR-17-25). ETS.
- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement*, 4, 105–126.
- Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, 17, 221–240.
- Kano, Y., & Azuma, Y. (2003). Use of SEM programs to precisely measure scale reliability. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics*. Springer. https://doi.org/10.1007/978-4-431-66996-8_14
- Karimi, L. (2015). *Model-based reliability and validity of measurement models using structural equation modeling* [Unpublishing doctoral dissertation, Swinburne University of Technology].
- Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, 22, 29–41.
- Kelley, T. L. (1947). *Fundamentals of statistics*. Harvard University Press.
- Kendall, M. G., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed., Vol. 1). Macmillan.

- Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, 58, 587–599.
- Kim, K. Y., & Lee, W. (2018). Confidence intervals for weighted composite scores under the compound binomial error model. *Journal of Educational Measurement*, 55, 152–172.
- Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, 77, 153–162.
- Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, 11, 179–188.
- Kim, S., & Livingston, S. A. (2017). *Accuracy of a classical test theory-based procedures for estimating the reliability of a multistage test* (ETS Research Report No. RR-17-02). ETS.
- Kim, S. Y., & Lee, W. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education*, 32, 97–115.
- Kim, S. Y., & Lee, W. (2020). Classification consistency and accuracy with atypical score distributions. *Journal of Educational Measurement*, 57, 286–310.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer.
- Kolen, M. J., & Harris, D. J. (1987, April 20–24). *A multivariate test theory model based on item response theory and generalizability theory* [Paper presentation]. American Educational Research Association Annual Meeting, Washington, DC, United States.
- Kolen, M. J., & Lee, W. (2011). Psychometric properties of raw and scale scores on mixed-format tests. *Educational Measurement: Issues and Practice*, 30, 15–24.
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29, 285–307.
- Kolen, M. J., Wang, T., & Lee, W. (2012). Conditional standard errors of measurement for composite scores using IRT. *International Journal of Testing*, 12, 1–20.
- Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33, 129–140.
- Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28, 221–238.
- Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, 39, 491–499.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Lawrence, I. M. (1995) *Estimating reliability for tests composed of item sets* (ETS Research Report No. RR-95-18). ETS.
- Lee, W. (2005). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Center for Advanced Studies in Measurement and Assessment, University of Iowa. <http://www.education.uiowa.edu/casma>
- Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement*, 31, 255–274.

- Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47, 1–17.
- Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement*, 37, 1–20.
- Lee, W., Brennan, R. L., & Kolen, M. J. (2006). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics*, 31, 261–281.
- Lee, W., Brennan, R. L., & Wan, L. (2009). Classification consistency and accuracy for complex assessments under the compound multinomial model. *Applied Psychological Measurement*, 33, 374–390.
- Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, 26, 412–432.
- Lee, W., Harris, D. J., Liu, H., & Chang, K. (2025). *Empirical investigation of various reliability statistics* (CASMA Research Report No. 58). Center for Advanced Studies in Measurement and Assessment, University of Iowa.
- Lee, W., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement*, 57, 230–254.
- Linacre, J. M. (1989). *Multi-faceted Rasch measurement*. MESA Press.
- Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Houghton-Mifflin.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9, 13–26.
- Livingston, S. A. (2018). *Test reliability—basic concepts* (ETS Research Memorandum No. RM-18-01). ETS.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, 19, 171–200.
- Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325–336.
- Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, 17, 510–521.
- Lord, F. M. (1965). A strong true score theory, with applications. *Psychometrika*, 30, 239–270.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233–245.
- Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, 21, 239–243.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Measurement in Education*, 8, 452–461.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity of examination scores. *Applied Measurement in Education*, 3, 331–345.
- Madison, M. J. (2019). Reliability assessing growth with longitudinal diagnostic classification models. *Educational Measurement: Issues and Practice*, 38, 68–78.
- Marcoulides, K. M. (2019). Reliability estimation in longitudinal studies using latent growth curve modeling. *Measurement: Interdisciplinary Research and Perspectives*, 17(2), 67–77. <https://doi.org/10.1080/15366367.2018.1522169>
- Margolis, M. J., & Feinberg, R. A. (2020). *Integrating timing considerations to improve testing practices*. Routledge.
- Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika*, 14, 189–229.
- Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. *Applied Psychological Measurement*, 39, 314–329.
- Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, 23(2), 351–362. <https://doi.org/10.1037/met0000132>
- Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, 23, 239–247.
- Oberski, D. L., & Satorra, A. (2013). Measurement error models with uncertainty about the error variance. *Structural Equation Modeling: A Multidisciplinary Journal*, 20, 409–428. <https://doi.org/10.1080/10705511.2013.797820>
- Ogasawara, H. (2009). Stratified coefficients of reliability and their sampling behavior under nonnormality. *Behaviormetrika*, 36, 49–73.
- Park, R., Kim, J., Chung, H., & Dodd, B. G. (2017). The development of MST test information for the prediction of test performance. *Educational and Psychological Measurement*, 77, 570–586.
- Patterson, C. H. (1955). The interpretation of the standard error of measurement. *The Journal of Experimental Education*, 23, 247–252.
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27, 341–384.
- Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement*, 17, 359–368.
- Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer administered measure of vocabulary. *Reading and Writing*, 28, 31–56. <https://doi.org/10.1007/s11145-014-9518-z> PMC 5053774

- Pugh, R. C., & Brunza, J. J. (1975). Effects of a confidence-weighted scoring system on measures of test reliability and validity. *Educational and Psychological Measurement*, 35, 73–78.
- Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, 30, 39–56.
- Raju, N. S. (1970). New formula for estimating total test reliability from parts of unequal length. *Proceedings of the 78th Annual Convention of the American Psychological Association*, 5, 143–144.
- Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika*, 42, 549–565.
- Raudenbush, S. W., & Jean, M. (2012). *How should educators interpret value-added scores?* http://www.carnegieknowledge.network.org/wp-content/uploads/2012/10/CKN_2012-10_Raudenbush.pdf
- Raykov, T., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable analysis. *Structural Equation Modeling*, 17, 629–641. <https://doi.org/10.1080/10705511.2010.510052>
- Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, 9, 195–212. https://doi.org/10.1207/S15328007SEM0902_3
- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω : A tutorial. *Psychological Assessment*, 31(12), 1395–1411. <https://doi.org/10.1037/pas0000754>
- Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, 16, 157–252.
- Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, 20(4), 335–343. <https://doi.org/10.1111/j.1745-3984.1983.tb00211.x>
- Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(14).
- Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13).
- Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*, 9, 99–103.
- Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion reliability: Implications for validation research. *Personnel Psychology*, 55, 807–825.
- Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, 58, 119–138.
- Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika*, 58, 195–209.
- Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, 18, 229–244.

- Sánchez-Meca, J., López-Pina, J. A., Rubio-Aparicio, M., Marín-Martínez, F., Núñez-Núñez, R. M., López-García, J. J., & López-López, J. A. (2019, May 30). *REGEMA: Guidelines for conducting and reporting reliability generalization meta-analyses*. Leibniz Institute for Psychology Information.
- Schmitt, T. A., Sass, D. A., Sullivan, J. R., & Walker, C. M. (2010). A Monte Carlo simulation investigating the validity and reliability of ability estimation in item response theory with speeded computer adaptive tests. *International Journal of Testing*, 10, 230–261. <https://doi.org/10.1080/15305058.2010.488098>
- Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement*, 23, 347–362.
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, 24(1), 70–91. <https://www.researchgate.net/publication/327483449>
- Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79–97.
- Seo, D. G., & Jung, S. (2018). A comparison of three empirical reliability estimates for computerized adaptive testing (CAT) using a medical licensing examination. *Frontiers in Psychology*, 9, 681. <https://doi.org/10.3389/fpsyg.2018.00681>
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
- Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment*, 20, 46–65.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13, 265–276.
- Sun, A., Valiga, M. J., & Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education*, 65, 367–379.
- Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, 30, 251–275.
- Terry, L., & Kelly, K. (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology*, 65, 371–401.
- Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–184). Lawrence Erlbaum.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scores in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum.

- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, 19, 39–49.
- Thomas, R., & Zumbo, B. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement*, 72, 37–43.
- Thompson, W. J., Clark, A. K., & Nash, B. (2019). Measuring the reliability of diagnostic mastery classifications at multiple levels of reporting. *Applied Measurement in Education*, 32, 298–309. <https://doi.org/10.1080/08957347.2019.1660345>
- Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). American Council on Education.
- Tisak, J., & Tisak, M. S. (1996). Longitudinal models of reliability and validity: A latent curve approach. *Applied Psychological Measurement*, 20(3), 275–288. <https://journals.sagepub.com/doi/abs/10.1177/014662169602000307>
- Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16, 8–14.
- Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalizations: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, 62, 562–569.
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46, 247–272.
- Verhelst, N., & Verstralen, H. (2001). IRT models for multiple raters. In A. Boomsma, M. van Duijn, & T. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). Springer.
- Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 16, 22–29.
- Wainer, H., Wang, X. A., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPoP curve. *Journal of Educational Measurement*, 42, 272–281.
- Walsh, P., Thornton, J., Asato, J., Walker, N., McCoy, G., Baal, J., Baal, J., Mendoza, N., & Banimahd, F. (2014). Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ*, 2, e651. <https://doi.org/10.7717/peerj.651>
- Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). University of Iowa.
- Wang, C. (2014) Improving measurement precision of hierarchical latent traits using adaptive testing. *Journal of Educational and Behavioral Statistics*, 39, 452–477.
- Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement*, 37, 141–162.
- Wang, Z., & Yao, L. (2013). *The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items* (ETS Research Report. No. RR-13-23). ETS.

- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.
- Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement*, 58, 21–37.
- Wood, S. W. (2020). Public perception and communication around automated essay scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 133–150). Chapman & Hall/CRC.
- Woodruff, D. J. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement*, 25, 191–208.
- Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement*, 13, 33–43.
- Woodruff, D. J., Traynor, A., Cui, Z., & Fang, Y. (2013). *A comparison of three methods for computing scale score conditional standard errors of measurement* (ACT Research Report Series 20137). ACT.
- Woodruff, D. J., & Wu, Y.-F. (2012). *Statistical considerations in choosing a test reliability coefficient* (ACT Research Report Series 2012, 10). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-10.pdf
- Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, 36, 602–624.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). American Council on Education/Praeger.
- Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data—which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16, 93. <https://www.frontiersin.org/articles/10.3389/feduc.2017.00049/full>
- Zhang, Y., Breithaupt, K., Tessema, A., & Chuah, D. (2006, April 8–10). *Empirical vs. expected IRT-based reliability estimation in computerized multistage testing* [Paper presentation]. A National Council of Measurement in Education Annual Meeting, San Francisco, CA, United States.

NOTES

1. Coefficient alpha can also be derived under the assumption of randomly parallel forms. It can be shown that alpha is equal to the generalizability coefficient of a $p \times I$ design in GT.
2. Refer to Brennan and Lee (2006) for some other perspectives on KR21, especially its relation to absolute error variance.
3. General procedures are available for the UAO consisting of a mixture of infinite and finite universes—that is, a mixed effects model. However, a univariate mixed model is often viewed as a special case of a multivariate model, in which there is a random effects model within each condition of a fixed facet.

Fairness in Educational Measurement

Theory and Concepts

Rebecca Zwick

ETS (Retired), University of California (Emerita)

Nearly 50 years after the publication of Lee Cronbach's landmark paper *Five Decades of Public Controversy Over Mental Testing* (1975) and a quarter century after Robert Linn's *A Century of Standardized Testing* (2001), test fairness continues to be very much in the public eye. The coronavirus pandemic that began in 2020 led to widespread cancellation of accountability testing in U.S. public schools and impelled many undergraduate and graduate institutions to eliminate admissions test requirements, in some cases permanently. A parallel development has been the initiation of large-scale remote testing programs, in which test takers respond to assessments in their homes or workplaces, often with human or electronic monitoring. These changes have been accompanied by a wholesale reconsideration of the fairness of standardized tests, a re-examination that has been further fueled by an increasing awareness of systemic racism in U.S. society and the role that test scores can have in limiting educational and occupational opportunities. Recent challenges to standardized testing have focused on the use of achievement tests as part of K–12 accountability programs, the fairness of college and graduate school admissions tests, and, more generally, the pervasive performance differences among socioeconomic and ethnic groups.

In striving to understand test fairness issues in the early 21st century, it is useful to briefly examine some key developments in the history of assessment. Although we tend to think of institutionalized testing programs and their attendant fairness and validity concerns as purely modern-day phenomena, the civil service examinations administered in imperial China provide a vivid counterexample. To qualify for a position, candidates had to take several rounds of exams in history, philosophy, painting, calligraphy, and poetry. Competition for the highest posts took place in Beijing every 3 years. Because these tests were the key to income and prestige, candidates' families often invested in years of special tutoring. Determined to pass at any cost, some candidates hired substitute test takers or purchased crib sheets ahead of time; others cheated by wearing special garments inscribed with prepared essays or by smuggling miniature books into the testing venue (Miyazaki, 1976; Webber, 1989).

Although the imperial civil service examination program began nearly 1,500 years ago (Miyazaki, 1976),¹ many of the evident fairness and validity issues are essentially the same as those we encounter today—those of access, privilege, and opportunity to learn. The tests were viewed as a step toward equity because government positions that had previously been filled by members of the aristocracy were now filled by individuals who successfully completed examinations that were considered impartial (Miyazaki, 1976). However, although the tests were theoretically open to all candidates, this was not the case in practice. First, “all candidates” actually meant “all men.” Second, only the upper classes had the resources and leisure time needed to become literate and master the topics of the exam. The wealthier classes were also more able to pay for intensive tutoring for candidates and for tools used to cheat—issues that have a particular resonance today. In addition, traveling to the testing site could be particularly difficult for candidates living in isolated areas, far from the capital. Compounding this problem was the fact that passing the tests was the only path to civil service positions. There

were no alternative opportunities for candidates to demonstrate their competence. Contemporary educational and occupational tests are often challenged on similar grounds.

The imperial civil service tests ended in the early 20th century. During their lengthy existence, various other types of institutionalized testing began to develop, including college admissions testing. The earliest college admissions test may have been a Sorbonne entrance examination administered in the 13th century (Office of Technology Assessment, 1992), although there is some debate on this point. Most historians agree that in England and Germany, admissions tests had been instituted by the mid-19th century. In most European countries, university graduation tests predated admissions tests. In the early 19th century, for example, anyone who could afford it could attend Oxford or Cambridge. Whether they could graduate was determined by examination (Webber, 1989). A milestone in the development of college admissions tests in the United States was the formation in 1900 of the College Entrance Examination Board (now College Board) by the leaders of 12 northeastern colleges and universities. The board created a set of entrance examinations that were administered by the member institutions and then sent back to the board for scoring.

In 1904, the same year that the last imperial civil service exam took place (Miyazaki, 1976) and shortly after the formation of the College Board, Alfred Binet, director of the psychology laboratory at the Sorbonne, developed a set of tasks to measure reasoning skills. Having abandoned skull measurements and other “medical” methods for assessing intelligence, he was now determined to use a “psychological” approach (Gould, 1996). The tasks were ordered, and each was assigned an age level. Binet was opposed to strict hereditarian theories of intelligence. From his perspective, intelligence was not fixed; it could be augmented through education. The primary purpose of intelligence testing was to identify and help children who needed special education (Gould, 1996). Binet’s tasks later formed the basis for the Stanford–Binet intelligence test (Terman, 1916), which came to be interpreted and used in ways that Binet would not have supported.

A staunch proponent of intelligence tests was Carl Brigham, a Princeton University psychology professor who was firmly entrenched in the eugenics movement. Based on his assessments of World War I military personnel, Brigham concluded that immigrants, Black people, and Jews were deficient in intelligence and that their presence could lead to a deterioration of the nation’s intellectual capacity (Brigham, 1923). Although he later acknowledged that existing tests were not adequate for making comparisons of the kind that he had conducted (Brigham, 1930, p. 165), he did not retract his claims about group differences in intelligence (Saretzky, 1982). As an advisor to the College Entrance Examination Board, Brigham chaired the commission that developed the multiple-choice exam that was eventually dubbed the Scholastic Aptitude Test, first administered in 1926. Although today’s SAT (no longer an acronym) bears little resemblance to the 1926 version, the SAT’s ancestral link to intelligence tests and eugenics continues to be cited in present-day criticisms of the test (e.g., Atkinson, 2001; see also Camara, 2009).

During the early 20th century, standardized tests gradually made their way into the public school system in the United States. Assessments in the schools proceeded on two fronts—intelligence testing and achievement testing. The National Intelligence Tests, a group assessment intended for schoolchildren, was introduced in final form in 1920 (Office of Technology Assessment, 1992). The assessment was an adaptation of the Army Alpha, a test used for screening military personnel during World War I (Whipple, 1921). The Army Alpha, which formed the basis of Brigham's best-known work, *A Study of American Intelligence* (1923), had itself been influenced by early intelligence tests. To measure student achievement, the psychologist E. L. Thorndike and his students developed standardized tests in arithmetic, reading, and other areas. By 1918, more than 100 standardized achievement tests, created by different researchers, were available to measure the skills of elementary and secondary school students (Office of Technology Assessment, 1992).

More than a century later, testing in U.S. classrooms has expanded tremendously. The federal Every Student Succeeds Act, approved in 2015, requires that states must “establish college- and career ready standards and maintain high expectations when assessing all students against those standards” (U.S. Department of Education, 2017). Currently, classroom tests are used to evaluate teachers, principals, and schools, as well as students (see Brookhart & DePascale, this volume, for discussion of classroom tests, and Ho & Polikoff, this volume, for discussion of accountability tests).

Other tests that are widely used today are those required for admission to various levels of schooling, graduation tests, employment tests, licensing and certification tests, language tests, military personnel tests, large-scale surveys such as the National Assessment of Educational Progress and the Programme for International Student Assessment, and personality tests. Many of these types of assessments are discussed in later sections of this chapter.

CHAPTER OVERVIEW

The remaining five sections of this chapter are organized as follows. The section “Concepts of Fairness” begins with a consideration of the relationships among the concepts of test fairness, validity, and bias. The section also describes the fairness perspectives represented in the *Standards for Educational and Psychological Testing* (the *Standards*; American Educational Research Association [AERA] et al., 2014), addresses the misguided notion of fairness as equality of performance, and discusses aspects of test content and test administration that have implications for fairness. The key principle of opportunity to learn is presented as well. The section “Legal Perspectives on Test Fairness” highlights four landmark court cases involving test construction practices, test disclosure policies, the annotation of test scores resulting from nonstandard administrations, and the equitability of college admissions tests, respectively. The section “Statistical Perspectives on Test Fairness: Measurement, Prediction, and Linking Invariance” explains how investigations of invariance contribute to the assessment of

test fairness. To permit fair test-based inferences, key aspects of test functioning should be invariant across demographic groups. This section describes the types of statistical analyses currently used in these evaluations, including assessments of differential item functioning, differential prediction, and differential validity. Examples of invariance studies are provided. The section “Selection Fairness” describes various models that have been proposed to improve the fairness of selection procedures. The chapter ends with the “Summary and Conclusion.”

Elsewhere in this volume, Rodriguez and Thurlow discuss fair and valid test score interpretations and uses, focusing on issues of accessibility for all test takers, including individuals with disabilities and English learners. Other recent treatments of the fairness of educational tests include Zwick and Dorans (2016), which considers philosophical perspectives on fairness; Zwick (2019a), which addresses selection fairness; Sireci and Randall (2021), which describes the evolving notions of test fairness in the United States; and Zwick (2021), which addresses the fairness of IQ testing, admissions testing, and accountability testing, including the use of value-added modeling to assess teachers. Fairness issues in the assessment of personal characteristics (sometimes called noncognitive factors) are addressed by Kyllonen and Zu (this volume).

CONCEPTS OF FAIRNESS

This section first considers the interrelated concepts of validity, fairness, and bias. Next, the section describes the concepts of fairness as presented in the *Standards* (AERA et al., 2014) and then addresses the perception that fairness implies equality of performance. Subsequently, fairness issues in test content are considered. The issue of opportunity to learn is also discussed in this section. Next, fairness considerations involving test administration practices are described, including issues arising in remotely proctored testing, a mode of test administration that became widespread during the coronavirus epidemic. Finally, the importance of treating test takers respectfully is addressed.

Test Validity, Fairness, and Bias

According to current perspectives in the field of educational measurement, validity refers to the degree to which a particular use or interpretation of an assessment is supported by evidence. If a test is unfair in a certain context—for example, a math test in English that is found to underestimate the math skills of native Spanish speakers relative to those of native English speakers—it is, by definition, invalid for the population and purpose in question: The test cannot be counted on to provide an accurate conclusion about the math proficiency of native Spanish speakers. When we evaluate the fairness of tests for particular groups, we are, of course, making generalizations. In this example, we may declare the test to be unfair to native Spanish speakers even though we realize that some native Spanish speakers have excellent English skills and some native English speakers do not. In this example, the precise meaning of the declaration of unfairness is that (a) English language skills are not part of the construct that the math

test is intended to measure and (b), on average, the English language skills of native Spanish speakers are not as strong as those of native English speakers (given a particular level of math proficiency). (See Lane & Marion, this volume, for a further discussion of validity and validation.)

Clearly, fairness and validity are inextricably connected. As stated in the *Standards* (discussed further in the section “Representation of Fairness in the *Standards for Educational and Psychological Testing*”), “fairness is a fundamental validity issue” (AERA et al., 2014, p. 49).

Bias is another term that arises in discussions of test fairness and validity. The term lacks a clear and precise definition in the measurement setting, partly because its meanings in other contexts have led to confusion. In mathematical statistics, a biased estimator is a statistic whose *expectation* is not equal to the target parameter value.² In everyday speech, bias means “prejudice” or “partiality.” In the world of educational measurement, *bias* has been assigned at least two different meanings, which are distinct from the statistical and everyday definitions. As described in the section “Prediction Invariance,” Cleary (1968) defined test bias as a phenomenon that can occur when a test score is used to predict a criterion measure—say, subsequent grades—using regression analysis: If predictions are systematically too high or too low for some group of test takers, this constitutes test bias. (This phenomenon later came to be called *predictive bias*.) Cole and Moss (1989) defined bias much more broadly, as “differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers” (p. 205). The hypothetical math test described earlier would be a biased test of the math abilities of native Spanish speakers, according to Cole and Moss (1989). Whether it was biased in Cleary’s sense would need to be determined by further analyses.

In practice, most fairness evaluations depend on the intended use of the test results and on the characteristics of the test takers. In general, then, fairness needs to be assessed with reference to a purpose and a test-taking population. This chapter, however, does not include separate sections that discuss fairness findings for particular populations, such as women, people of color, English learners, and students with disabilities. Instead, examples involving various groups are interwoven throughout the chapter.

Validity, Fairness, and Test Consequences

In the late 1970s, two key legal cases about the consequences of educational testing were argued: *Larry P. v. Riles*, which involved the use of IQ tests to assign children to special education classes, a practice that resulted in the disproportionate representation of Black students; and *Debra P. v. Turlington* (see the section “Legal Perspectives on Test Fairness”), which concerned the use of graduation tests that were based on materials that some students had not had the opportunity to learn. These court cases brought to the fore the stark consequences that can result from testing policies.

The question of whether the social consequences of testing need be evaluated as part of a validity assessment became a prominent topic of debate that flourished in the 1970s and 1980s and has continued to the present day. These arguments are discussed in some

detail by Messick (1989), Kane (2006), Kane and Bridgeman (2021), and Lane and Marion (this volume). Messick (1989) explicitly included the social consequences of test use as a facet of validity:

The interpretation of test scores and . . . the use of test scores in the implementation of social policy falls well within the realm of validity inquiry, because the import of scores for action depends on the validity of their meaning and their value implications. (p. 63)

Cronbach, like Messick, argued for the importance of considering testing consequences (e.g., Cronbach, 1988), although their stances differed somewhat. Messick argued that negative consequences of testing “counted against” validity when they resulted from flaws in the test. Cronbach, however, argued that negative consequences were important to consider regardless of their cause, while remarking that this determination might not be part of the validation process (see Kane & Bridgeman, 2021). In subsequent contributions to the literature, Linn (1997), Moss (1998), and Shepard (1997) supported including the evaluation of consequences in test validation, while other measurement specialists, such as Mehrens (1997) and Popham (1997), argued for a narrower definition of validity assessment.

Once we acknowledge that fairness assessment is part of validation, however, there can be no debate as to the relevance of values and social consequences: These are central to the concept of fairness. In noting that some validation inquiries are “too narrow” because of their failure to consider social consequences, Cronbach (1988) highlighted an issue that particularly resonates today:

The obvious example of too narrow an inquiry is the validation of employment tests, which used to concentrate on predicting a criterion the employer cared about and neglected to collect facts on what kinds of applicants were most likely to be rejected by the test. (p. 7)

As this example makes clear, neither validity nor fairness can be unambiguously defined via technical means: Values are inevitably involved. In the situation mentioned by Cronbach (1988), a decision would be needed as to what demographic groups of employment candidates should be studied in evaluating test consequences (see the section, “Statistical Perspectives on Test Fairness”) and what outcomes would be acceptable.

To summarize, validity and fairness are inextricably bound together. A test-based inference cannot be valid without also being fair. Social values and consequences must be considered in evaluating fairness and thus are essential to determining validity as well.³

Representation of Fairness in the *Standards for Educational and Psychological Testing*

The *Standards*, issued jointly by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education, provides “criteria for the development and evaluation of tests and testing

practices and . . . guidelines for assessing the validity of interpretations of test scores for the intended test uses” (AERA et al., 2014, p. 1). As noted in the document itself, “it has been repeatedly recognized by regulatory authorities and courts as setting forth the generally accepted professional standards that developers and users of tests and other selection procedures follow” (p. 2).

Fairness was not explicitly discussed in the first two editions of the test standards, which appeared in 1954 and 1966, respectively (Sireci & Randall, 2021). During the civil rights era, issues of test bias and group performance differences became a more prominent feature of test validity discussions, and the 1974 *Standards* (American Psychological Association et al., 1974) included considerations of these issues (Messick, 1989), as did the 1985 and 1999 *Standards*.

The most recent edition, published in 2014, is the sixth version of this publication. In that edition, fairness is included in the “Foundations” section, which also contains the chapters on validity and reliability. The prominence of the fairness material reflects an attempt by the committee in charge of the revision to broaden the perspective on fairness and elevate its importance, relative to earlier editions. The committee made the decision not to include individual chapters on particular subpopulations in order to extend “the scope of fairness to include all examinees in the intended population of test takers” (Plake & Wise, 2014, p. 7; see Sireci & Randall, 2021, for further discussion of the treatment of fairness in the *Standards*).

The volume includes 20 fairness standards, beginning with an overarching standard stating that “all steps in the testing process . . . should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations” (AERA et al., 2014, p. 63).⁴ Construct-irrelevant variance is variability in the scores of a group of test takers that is attributable to characteristics, knowledge, or skills that are distinct from the construct the test is intended to measure.

The remaining 19 fairness standards are organized into four thematic clusters (AERA et al., 2014, p. 63):

1. Test design, development, administration, and scoring procedures that minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups
2. Validity of test score interpretations for intended uses for the intended examinee population
3. Accommodations to remove construct-irrelevant barriers and support valid interpretations of scores for their intended uses
4. Safeguards against inappropriate score interpretations for intended uses

The key principles presented in each of these clusters are briefly described here, followed by a discussion of the fairness perspective adopted in the *Standards*.

Cluster 1 notes that test developers must clearly define the construct that is to be measured, along with the characteristics of the intended population (Standard 3.1). In addition, the language used in tests should be appropriate to the purpose of

the test and should be familiar to “as wide a range of test takers as possible” (3.2). (The familiarity requirement does not apply to language that is a legitimate aspect of the construct that is being tested, as in a vocabulary test.) Also, “relevant subgroups” should be included in technical studies of tests and reviews of test material (3.3). An important concept described in this cluster is *universal design* (3.1), a test development process that “emphasizes the need to develop tests that are as usable as possible for all test takers in the intended test population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability” (AERA et al., 2014, p. 57). (Further discussion of universal design appears in the section “Fairness of Test Content.”)

Cluster 2 asserts the importance of investigating reasons for subgroup mean differences, with an eye toward determining whether construct-irrelevant factors could be responsible (3.6), notes that studies of differential prediction should be conducted where relevant and possible (3.7; see the section “Prediction Invariance” for further discussion), and states that the possibility of biases in human and automated scoring of constructed responses should be investigated (3.8).

Cluster 3 describes the importance of “providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers” that otherwise would compromise test takers’ performance (3.9) and the need to monitor and document implementation of any accommodations (3.10) or translations (3.12). When tests are altered to improve accessibility or translated from one language to another, studies should be undertaken to examine the efficacy of the accommodation or translation and to determine whether scores are comparable to those obtained from the original test (3.12, 3.13).

The standards of Cluster 4 state that tests should be used only for subgroups for which sufficient validity evidence exists to support the desired score interpretations while acknowledging that this principle may be superseded by legal requirements in some cases (3.16). When test scores are reported in aggregate form for multiple groups of test takers, evidence should be provided that scores have comparable interpretations across groups. If evidence of this kind is not available, appropriate cautionary statements should be attached to the reports (3.17). In diagnostic settings, test users should not use test scores as the sole measure of an individual’s capabilities; additional factors such as age, culture, and language should be considered (3.18). A key standard appearing in this cluster concerns opportunity to learn (see also the section “Opportunity to Learn”). According to Standard 3.19, when

the same authority [such as a state] is responsible for both provision of curriculum and high-stakes decisions based on testing . . . curriculum mastery, examinees should not suffer permanent negative consequences if . . . they have not had the opportunity to learn the test content. (p. 72)

Despite the expanded treatment of fairness in the 2014 *Standards*, the general perspective—that “the central idea of fairness in testing is to identify and remove

construct-irrelevant barriers to maximal performance” (AERA et al., 2014, p. 63)—has been criticized as being too narrow (Lorrie Shepard, personal communication, February 25, 2019). Shepard, a pioneer in the area of test fairness, takes a broad view of fairness issues, asserting that “equity in assessment requires a validity framework that evaluates whether an assessment system produces its intended positive effects without incurring familiar negative effects on students’ opportunities to learn” (Shepard, 2012, p. 812). She adds that “despite advances in statistical checks and test-making strategies, . . . the more fundamental question persists about when it is appropriate to use fallible tests that affect students’ life chances” (p. 809). (This issue is further considered in the “Summary and Conclusion” section.)

In this chapter, I consider fairness in testing and selection from a somewhat broader perspective than that taken in the *Standards*, devoting more attention to opportunity to learn, selection processes, and alternatives to testing. The *Standards* remains a key resource on test fairness that will continue to be recognized as a representation of the accepted standards among members of the measurement community.

Fairness as Equality of Performance

Outside the measurement community, a common belief is that a fair test will produce equal average test scores for all recognized test-taker groups (Zwick, 2002, 2017a). If average scores are unequal, this is often seen as a sufficient demonstration that the test is unfair. As noted by Zwick (2017a, p. 35), “endorsement of the equal performance principle often takes the form of automatically interpreting any group differences in test scores as a manifestation of unfairness or bias.” Although this view is explicitly rejected by the *Standards* (AERA et al., 2014, p. 54), it has been embodied in court settlements and proposed legislation and featured in countless news articles and testing critiques. An example is a bill introduced in the California Senate in 1998 and 1999 that stated that “a test discriminates . . . if there is a statistically significant difference in the outcome on test performance when test subjects are compared on the basis of gender, ethnicity, race, or economic status” (Hughes, 1998). If the bill had passed, test sponsors could have been fined \$750 per test taker for administering a test that showed performance differences among student groups.

Support for this perspective on fairness stems from well-founded concerns about the pervasive finding of large differences among demographic groups in average scores on standardized tests. On educational assessments, students from high-income families typically score higher, on average, than students from lower-income families, and on many tests, Asian and White students tend to score higher than Black, Latino, and Native American students. Gender differences are found on some tests as well. Clearly, these patterns warrant investigation, including examination of the test’s content and administration procedures for possible sources of bias.

In investigating these score differences, however, it is important to consider that other educational achievement measures often show patterns of performance similar

to those observed for tests. For example, high school grade point averages (GPAs) and college admissions test scores show similar patterns of results across ethnic and socioeconomic groups (e.g., Camara & Schmidt, 1999; Owings et al., 1995). Results from the Education Longitudinal Study, based on a nationally representative sample of the high school seniors of 2004, show that on admissions tests, the difference in average scores between the highest- and lowest-scoring ethnic groups was 1.3 standard deviation units (similar to recent ACT and SAT results) and the difference between the highest and lowest socioeconomic quartile was 1.1 units. The corresponding differences for high school GPA were 1.0 and 0.7. Within each ethnic group, both test score and GPA averages increased with socioeconomic level (Zwick, 2019b). (Technical reasons for the smaller differences on GPA than on test scores include reliability differences; see Zwick & Green, 2007.)

To conclude that these differences in average test scores and grades are in themselves evidence of unfairness would be sensible only if educational opportunities were the same for all groups in question, allowing all students to develop the same level of academic skill. In fact, opportunities and resources vary widely across communities. For example, teacher experience, teacher credential status, and student:teacher ratios vary with family and school-level socioeconomic status (SES; Betts et al., 2000; Organisation for Economic Co-operation and Development, 2010). In addition, childhood poverty rates for Black, Hispanic, and American Indian children in the United States are much higher than those for Asian and White children (National Center for Education Statistics, 2022). Even in kindergarten, ethnic and socioeconomic groups differ in school performance (Lee & Burkam, 2002).

Given these inequities, it would be surprising if educational achievements were equivalent for all student groups. At the same time, the existence of many plausible explanations for group differences should never serve as a reason to assume that tests are blameless.

Fairness of Test Content

Validity theorists such as Messick (1989) and Kane (2006) have noted that content-related evidence has only a limited role to play in test validation. Kane states that evidence that the tasks included in the test are representative of the domain and that the scores are generalizable “can make a basic positive case for the validity of an interpretation in terms of expected performance over some universe of possible performances, but to go beyond that basic interpretation, other kinds of evidence are needed” (2006, p. 19). Just as content evidence alone cannot be sufficient to demonstrate validity, neither can it demonstrate fairness. An evaluation of test content, however, is a crucial first step in assessing test fairness. In fact, test content must be scrutinized throughout the test development process, not merely as part of a post hoc review process.

Basic fairness principles dictate that demographic groups must be referred to using currently accepted terms and must be portrayed in a balanced and respectful way in test questions and reading passages. Many of the principles of universal design

of assessments have implications for test content as well. According to the National Center on Educational Outcomes (2016), “the goal of universal design principles is to improve access to assessments for all students” (para. 1). Noting that “more accessible tests may provide a more accurate understanding of what students know and can do,” the organization developed a list of seven elements of universal design, the first of which is an “inclusive assessment population.” The remaining principles are relevant to test content: The test should be based on precisely defined constructs and should be amenable to accommodations. It should contain accessible, nonbiased items; should feature simple, clear, and intuitive instructions and procedures; and should be designed for maximum readability, comprehensibility, and legibility.

Additional aspects of content fairness are considered in the following sections. The issues of construct representation, sensitive or unnecessarily challenging material, and coachability and vulnerability to compromise are addressed in the sections on “Construct Representation,” “Sensitive or Unnecessarily Challenging Material,” and “Coachability and Vulnerability to Compromise,” respectively. Opportunity to learn—a central issue in certain achievement-based tests—is discussed in a separate section within this chapter.

Construct Representation

As noted in the section “Concepts of Fairness,” the *Standards* (AERA et al., 2014) focuses heavily on the importance of minimizing construct-irrelevant variance in test scores. Ideally, individuals’ scores should not differ as a result of factors that are distinct from the characteristic that the test is intended to measure. In a test meant to assess only computational skills, for example, it is undesirable for test takers’ scores to vary because some have difficulty understanding the vocabulary used in the questions. A key concern is that the inclusion of construct-irrelevant factors may disadvantage certain populations. In this case, for example, English learners could be particularly harmed by the inclusion of difficult vocabulary. Further examples of construct-irrelevant test features are described in the sections on “Sensitive or Unnecessarily Challenging Material” and “Fairness of Test Administration and Scoring Procedures.”

Construct underrepresentation, a phenomenon in which a test fails to include key components of the intended construct, receives far less attention in the *Standards* than construct-irrelevant variance, but can have substantial implications for fairness. Consider the case of an employment test for an administrative assistant position. Suppose that answering phone calls and preparing correspondence are equally important features of the work, but that only keyboarding and English-language writing skills are included in the test. Because a key aspect of job performance—communicating by phone—is omitted, the needed job skills are incompletely represented by the test material. The underrepresentation would be particularly disadvantageous to test takers who excelled at telephone communications but were less skilled at keyboarding because of a disability or were less proficient in writing because they were English learners.

In the case of academic assessments, a key content-related concept is *alignment*. Tests that are designed to assess whether students have learned content that has been specified in curriculum materials are expected to be aligned with that content. These types of exams, such as end-of-year assessments in K–12 education, are expected to sample the prescribed content and to exclude material that was not part of the curriculum. Of course, inclusion in curricular materials is no assurance that material has been taught in a satisfactory way. This issue is discussed further in the section “Opportunity to Learn.”

Sensitive or Unnecessarily Challenging Material

Fairness requires that the content and format of a test not be more challenging for the test taker than required for the test’s purpose. According to the *Standards* (AERA et al., 2014),

Test developers should use language . . . that is consistent with the purpose of the tests and that is familiar to as wide a range of test takers as possible. Avoiding the use of language that has different meanings or different connotations for relevant subgroups . . . will help ensure that test takers . . . are able to understand what is being asked of them and respond appropriately. (p. 64)

In particular, a standard principle of test design is that esoteric or specialized vocabulary, such as military jargon, regionalisms, or terminology specific to U.S. culture, be avoided unless it is important for valid measurement.

More debatable is the idea that, unless they are needed for valid measurement, disturbing or polarizing topics should be excluded as well. The *ETS Guidelines for Developing Fair Tests and Communications*, for example, notes that

construct-irrelevant emotional barriers to success arise when language, scenarios, or images cause strong emotions that may interfere with the ability of some groups of test takers to respond to an item. . . . Test takers may respond emotionally rather than logically to excessively controversial material. (ETS, 2022, p. 24)

Examples of the topics of concern are accidents, illnesses, or natural disasters; advocacy of contested causes or ideologies; death and dying; sexual behavior; and depictions of “situations that are associated with excessive spending on what some members of the test-taking population would consider luxuries” (p. 29). The evolution of human beings, too, is listed as a topic to avoid except on tests in content areas that require knowledge of evolution.

In her 2003 book *The Language Police*, education historian Diane Ravitch expressed concern about the reviews of educational tests and textbooks that are now routine:

The bias and sensitivity reviewers work with assumptions that have the inevitable effect of stripping away everything that is potentially thought-provoking and colorful from the texts that children encounter. These assumptions narrow what children are exposed to, at least on tests and in textbooks. (p. 8)

As suggested by these comments, it is worth considering that rigid adherence to sensitivity guidelines carries its own risks: Test materials could become so bland and uncontroversial that they fail to represent the domain of interest or fail to engage respondents. These situations, like the inclusion of inappropriate material, have the potential to compromise the validity of the results.

Arguing from a perspective entirely different from Ravitch's, Randall (2021) deplored the fact that context "is routinely scrubbed from assessment items in service to a purified construct representation" (p. 82). Contending that "when the context is not clear (or seemingly not present), the implied context, historically, has been whiteness" (p. 83), Randall recommends "reenvisioning construct definition/representation in a way that decenters whiteness and acknowledges the ways of knowing and understanding associated with minoritized individuals as construct relevant, and arguably, construct critical" (p. 89).

This perspective is consistent with earlier literature on culturally relevant teaching and learning (see Ladson-Billings, 1995). Gutiérrez and Rogoff (2003), for example, discussed the importance of individuals' linguistic and cultural-historical repertoires—their "ways of engaging in activities stemming from observing and otherwise participating in cultural practices" (p. 22). They note that "individuals' background experiences, together with their interests, may prepare them for knowing how to engage in particular forms of language and literacy activities [and] play their part in testing formats" (p. 22). In particular, the authors noted that "students who have participated in varying cultural traditions would differ in repertoires for . . . answering known-answer questions [or] analyzing word problems on the basis of counterfactual premises" (p. 22).

In response to recent critiques, the educational measurement field is now undergoing a reconsideration of the historical principles that recommend the minimization of contextual material that is not directly related to the purpose of the assessment.⁵

Coachability and Vulnerability to Compromise

The possibility that some but not all test takers have the opportunity to improve their scores through test preparation programs is frequently raised as a criticism of standardized tests, particularly admissions tests. Although coaching on college admissions tests produces only modest gains on average (see Zwick, 2017a), unequal access to coaching, especially the more expensive forms, constitutes a threat to fairness, compounding the advantages that accrue to students from high-income families, such as better schooling, a wider range of cultural activities outside school, more educated parents, and more favorable environments for completing homework. More broadly, coaching threatens test validity because it can inflate scores without promoting any generalizable gains in knowledge or skills. Although it may be impossible to create a test that is not coachable, the problem of differential access to coaching can be ameliorated. Many testing programs in the early 21st century offer free online test preparation materials. Further improvement could be achieved through the establishment of widely available and inexpensive test preparation programs.

In the case of admissions testing, events that took place in 2019 made it obvious that wealthy families have easier access not only to costly test preparation, but also to opportunities for altering their children's test results that are not available to lower-income families. In a scandal nicknamed Varsity Blues, federal prosecutors charged 50 individuals in a scheme to purchase admission to elite universities. Most of those charged were parents of the applicants (Medina et al., 2019). In some of these cases, hired test takers took admissions tests in place of the actual college applicants; in others, test administrators altered the candidates' test scores. *The Code of Fair Testing Practices in Education* includes in its list of responsibilities of test users the responsibility for "eliminating opportunities for test takers to obtain scores by fraudulent means" (Joint Committee on Testing Practices, 2004, p. 7). Robust test security measures can reduce the likelihood of cheating in standardized test administrations. Reducing the incentive to cheat on tests is a more difficult goal, especially when educational and occupational opportunities are at stake.

Opportunity to Learn

Opportunity to learn is a well-established fairness principle that first received substantial formal consideration in the landmark *Debra P. v. Turlington* (1979, 1984) lawsuit, filed in 1978. In 1976, the state of Florida instituted a high school graduation test called the State Student Assessment Test, Part II (SSAT-II), a multiple-choice assessment of basic math and communication skills. Students were allowed to take the test more than once. After three administrations, the failure rate was 20% for Black students compared to 2% for White students.

The *Debra P.* plaintiffs charged that the SSAT-II was racially biased, that it was administered without sufficient notice, and that it was designed to place Black students in remedial classes. The case led to an examination of the fairness and validity of the test and an investigation of the students' opportunity to learn the material. In particular, how did Florida's history of discrimination and segregation affect Black students' opportunities to learn? More generally, were Florida schools teaching the material included on the SSAT-II?

Through 1967, Florida maintained completely separate school systems for Black and White students, with segregation continuing even longer in some counties. Court transcripts noted that the facilities, curricula, and textbooks for the Black schools were clearly inferior. The court ruled that this "past purposeful discrimination" was perpetuated by the use of the SSAT-II to deny high school diplomas and enjoined the state from using the test for this purpose until 1983. The intention was both to provide for adequate notice to test takers and to make sure that they would have had the chance to complete 12 years of schooling in integrated schools.

In subsequent court proceedings, the state provided results of a study intended to show that Florida students were being taught the material on the test, offered evidence of the remedial support available to those who initially failed the SSAT-II, and presented data on the much-improved pass rates for the class of 1983. A federal appeals court ruled in 1984 that the test could be used to award diplomas.

The *Debra P.* case underscores the fact that an educational test could be well aligned with the intended content domain and could have adequate psychometric properties yet could be unfair for certain purposes because some test-taker groups did not have sufficient opportunity to learn the material. In addition to appearing in Cluster 4 of the *Standards'* fairness principles (see the section "Representation of Fairness in the *Standards for Educational and Psychological Testing*"), opportunity to learn is also central to Standard 12.8 in the section on educational testing and assessment: "When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test" (AERA et al., 2014, p. 197).

As the *Standards* makes clear, the requirement that test takers have had the opportunity to learn the tested material applies only to particular testing situations—those in which institutionally administered curriculum-based achievement tests are used to make high-stakes decisions about individual students, such as those involving promotion or graduation. The requirement does not apply to other types of assessments, such as licensing tests, employment tests, or diagnostic tests.

Although the opportunity-to-learn (OTL) principle is intended to protect test takers from unfair consequences of test results, Randall (2021) cautioned that, by providing an excuse for group differences in performance, the existence of OTL disparities can be invoked in ways that undermine test fairness rather than support it. She maintained that the OTL concept "has long been used to disregard arguments against the use of particular assessments that have had a historical legacy of privileging whiteness and further marginalizing all others" (p. 86). She argued that OTL should be investigated only after "the assessment itself has been thoroughly interrogated for white-supremacist content and antiracist content moved into its place" (p. 87). Randall's commentary is a reminder of the importance of continuously re-evaluating test fairness principles and incorporating a wide range of perspectives.

Fairness of Test Administration and Scoring Procedures

Even if the test content itself is appropriate to the purpose of the test, administration and scoring procedures may produce construct-irrelevant variance in test scores. These issues are considered in detail by Rodriguez and Thurlow (this volume); only a brief discussion is included here.

In nearly all testing situations, it is undesirable for scores to differ because some test takers are unfamiliar with the computers used to administer the assessment (e.g., see Zwick & Bridgeman, 2014), find the print too small in a test booklet, or are made physically uncomfortable because of the testing conditions. Time restrictions that are not needed for valid measurement may also be an impediment to fairness, particularly for students with certain disabilities, such as attention deficit/hyperactivity disorder. These are examples of test administration factors that can introduce construct-irrelevant variance in test scores.

It is important to recognize that fair testing does not imply uniformity of test content or administration procedures across test takers. An obvious example is the use of testing accommodations. In this context, the departure from uniformity is intended to enhance fairness. Mislevy (2018) elaborated on this principle by introducing the concept of *conditional fairness*, which addresses “principled departures” from uniformity:

Sometimes task situations . . . or evaluation procedures . . . can be adapted to students in principled ways . . . such that even though the procedures are less similar across examinees, the resulting *evidence* is more comparable. (p. 219)

Accommodations intended to reduce or remove construct-irrelevant sources of variance are commonly used for students with disabilities such as blindness, deafness, or attention deficit/hyperactivity disorder. For example, screen reader software that transmits text in the form of synthesized speech may be available to a blind test taker or an assistant may be engaged to read the test material aloud. For test takers with learning disabilities, additional testing time is often requested.

A challenge involved in developing appropriate accommodations is that the construct assessed by the accommodated version of a test may depart from the intended construct. For example, if an assistant provides an oral description of a graph on a math test to a blind test taker, the construct may be subtly altered. It may not be possible to improve accessibility of the test material without modifying the construct. This is an extraordinarily difficult issue to investigate because both the test and the population are different from those in a standard administration. Often, attempts to validate particular test accommodations involve an effort to show that scores on the accommodated test bear the same relationship to other factors (say, course grades) as scores on the unaccommodated version. Research in this area is not always possible, however, because of the sparsity of data for particular combinations of disabilities and accommodations.

The test scoring process can also be a source of construct-irrelevant variance. This can happen, for example, if individuals scoring a test intended to assess students’ abilities to create a coherent and engaging essay are influenced by the students’ handwriting quality. Of course, test scores should not be influenced by any knowledge scorers may have about a test taker’s personality, appearance, classroom behavior, or performance on other tests or activities. Automated scoring procedures that have been developed by maximizing the similarity between automated and human scores (as in the case of automated essay scoring; see Shermis et al., this volume) must be examined to ensure that they have not “inherited” any biases that were present in the human scoring process.

Respectful Treatment of Test Takers

Although considerations of test fairness often focus on equal treatment of population groups, fairness also encompasses some general principles that do not depend on group membership. To take an obvious example, it would be unfair for a test proctor to offer high scores for a price, and this would be true regardless of the test’s purpose or the demographic groups to which test takers belonged.

More generally, fairness in measurement and selection requires that test takers be treated respectfully and impartially and that they be provided with information about the assessment, its scoring criteria, and its intended use, as noted in the *Standards* (AERA et al., 2014, pp. 133–134). As indicated in the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004), “fairness also extends to the accurate reporting of individual and group test results” (p. 2). This same document, developed as a cooperative effort among several professional organizations, points out test users’ obligation to provide timely test results in a comprehensible form, protect scores from unauthorized release, and “avoid using a single test score as the sole determinant of decisions about test takers” (p. 9; see also AERA et al., 2014, p. 198).

The availability of sophisticated technical methods for assessing particular aspects of test fairness should not distract measurement professionals and test users from the importance of respectful treatment of test takers, including adherence to clearly stated procedures, and straightforward communication about the purposes of tests and the meaning of test scores.

Fairness of Remotely Proctored Testing

The COVID-19 pandemic that began in 2020 led to a significant reduction in the frequency of educational testing and major changes in the administration of many tests. Because most school systems in the United States were closed for at least part of the year (Map: Coronavirus and School Closures, 2020), the number of tests administered at the K–12 level was greatly reduced. In addition, the majority of undergraduate institutions (Jaschik, 2020a) and some graduate programs (Hu, 2020) made the submission of admissions test scores optional for applications filed in 2020, decreasing the demand for this type of test. A number of large testing programs, including those offering the Advanced Placement tests, Graduate Management Admission Test (GMAT), Graduate Record Examinations (GRE), Law School Admission Test (LSAT), Test of English as a Foreign Language (TOEFL), HiSET high school equivalency test, and the Praxis tests for those entering the teaching profession, offered candidates the option of taking tests at home (or in another location they selected, such as a workplace or a relative’s home), with remote proctoring.

In remotely proctored testing programs, candidates take the test via the Internet, while being monitored by a test proctoring organization. Typically, the proctoring service installs software on test takers’ computers, preventing the use of software that would facilitate cheating. In addition, test-taker behavior, including unusual movement, is tracked using artificial intelligence approaches. A human proctor may also remotely monitor test takers and their surroundings, using the computer’s audio and video functions, to ensure that test takers are not relying on any collaborators or materials in answering test questions. Video recordings of testing sessions may be made as well.

Although remote testing is a welcome opportunity for many test takers, it brings with it many significant questions about the validity and fairness of this mode of test

administration. To what degree is test accessibility affected by inadequate computer equipment or Internet access, especially for low-income test takers or those from underserved ethnic groups? Are test takers with disabilities able to obtain the accommodations they need for optimal test performance? Are monitoring procedures applied equitably across groups? To what degree are test takers able to evade the proctors and obtain answers from outside sources? And ultimately, how do these factors affect test performance?

Candidates' ability to gain access to the test is a fundamental concern. To be eligible for home testing, the candidate must have a computer that meets certain specifications and must have Internet access. In explaining its decision not to offer an at-home version of the SAT, the College Board cited "concerns that many students would not have access at home to the three hours of reliable Internet service that would be required" (Anderson, 2020). According to 2019 data from the Digest of Education Statistics, 16.8% of individuals of ages 3 and above do not have Internet access at home, and the percentage varies widely over ethnic and income groups. For example, about 15% of Asian and White people are without access, compared to 20% to 23% for those who are Black, Hispanic, or American Indian/Alaska Native. For those with annual incomes of \$150,000 or more, the percentage without access is 8.6, compared to 34% to 35% for those with incomes less than \$20,000 (National Center for Education Statistics, 2020; see also Herold, 2020; Langenfeld, 2020).

Even if candidates have computing facilities that are deemed adequate, those with older computers and weaker Internet access may still be at a disadvantage because of the greater likelihood of disruptions. Test takers who do not have access to a quiet, uncluttered place to take the test may also be unable to perform as well as they would in a test center.

Another significant fairness issue is the availability of testing accommodations for students with disabilities. Certain accommodations that are possible in test centers may not be available in remote testing situations. In a class-action suit against the College Board in connection with its remote version of the Advanced Placement tests, plaintiffs claimed that "questions remained about the availability and applicability of legally required accommodations for students with disabilities" (*Class Action Lawsuit*, 2020). Similar issues were at the core of a complaint filed in connection with *Smith v. Regents of the University of California* (2019), a University of California case involving admissions test use (see the section, *Smith v. Regents of the University of California*). In a press release, attorneys for the plaintiffs stated that

even under normal circumstances, students with disabilities are required to navigate a byzantine set of requirements just to access the tests with the accommodations they need. These students must undergo costly evaluations and, since [University of California] does not require test sites to accept students' accommodations, must search (often in vain) for their own testing locations—all to take tests that have limited, if any, validity for them. This process has become virtually impossible to complete during the pandemic. (*Students Sue to Enjoin*, 2020)

A related issue involves test takers with disabilities who may have conditions that lead to involuntary movements or may engage in repetitive movements to remain alert. Concerns have been raised that these activities could be interpreted as cheating and could lead the proctoring software or a human proctor to trigger an alarm or lock the exam (Patil & Bromwich, 2020).

Another potential problem is that human proctors may apply monitoring procedures in ways that depend on the ethnicity or gender of the test takers. Biases of this kind have been observed in the classroom (The Graide Network, 2018; Rauschenberg, 2014; van Ewijk, 2011) and, of course, in noneducational settings, such as traffic stops. In addition, test-taker reports suggest that proctoring software may sometimes be unable to verify the identity of people of color, making it difficult for them to access their exams (Patil & Bromwich, 2020).

In addition, remote proctoring can be distracting and anxiety producing. Some test takers may be particularly disturbed by the loss of privacy (Langenfeld, 2020; Patil & Bromwich, 2020). A complaint filed in 2020 against five proctoring companies stated that “forcibly collecting personal information from test takers, including sensitive biometric data, is inherently invasive” (Chin, 2020), and in 2022, a federal court sided with a university student who argued that the scanning of his bedroom that was required as part of the remote proctoring of an exam was an unconstitutional search (Swaak, 2022).

Yet another issue is test security. Cheating on remotely administered tests may involve the purchase of devices that allow test takers to obtain assistance without alerting the proctoring system or the purchase of test items themselves. Like all test cheating methods, these workarounds are likely to be more easily available to wealthier test takers. Their use compromises the validity and fairness of the tests.

Despite these many challenges, eliminating the opportunity to take tests remotely would certainly be undesirable. Individuals who were required to take tests for admission to educational institutions, employment, professional licensing or certification, or course credit would be severely disadvantaged if these tests were unavailable in any form. A substantial research effort is needed to investigate and address the intertwined issues of technical adequacy, fairness, and privacy that affect remote testing.

LEGAL PERSPECTIVES ON TEST FAIRNESS

Controversies about the fairness of educational tests have frequently landed in the courtroom, sometimes resulting in significant changes in the content and use of tests. Rebell (1989) noted that “courts have had a significant impact on the formulation of public policy issues associated with testing, so much so that it may not be an exaggeration to say that they have become a major public policy forum in this area” (p. 135). Here, I describe four cases, each of which illustrates a key fairness issue: *Golden Rule Insurance Company v. Washburn*, *Association of American Medical Colleges v. Carey*, *Breimhorst v. Educational Testing Service*, and *Smith v. Regents of University of California*. (In addition, *Debra P. v. Turlington*, the landmark case involving opportunity to learn, is described in

the section on “Opportunity to Learn.”) Rebell (1989) and Phillips (2016) provided accounts of many other significant cases.

Golden Rule Insurance Company v. Washburn (1984)

In 1976, the Golden Rule Insurance Company sued the Illinois Department of Insurance in Illinois Circuit Court, charging that a set of employment tests for insurance agents, developed by Educational Testing Service for the department, was biased against African American test takers. The evidence presented was that African American test takers received lower scores on average, making it difficult for them to become licensed insurance agents. The case was notable in that its eventual out-of-court settlement, which occurred in 1984, included a specific set of compulsory procedures for test construction. In selecting items for use in a test form, preference was to be given to those for which the rates of correct response for Black test takers, White test takers, and all test takers combined were at least 40% and for which the rates of correct response for Black and White test takers differed by no more than 15 percentage points. (Formal hypothesis tests were to be used to determine whether these criteria were met.) This approach to test development came to be known as the Golden Rule procedure, a name that unfortunately seemed to suggest that the method was an equitable one. Subsequent court decisions and legislative proposals involving other assessments, such as teacher certification tests, embodied versions of the procedure (Faggen, 1987; Linn & Drasgow, 1987).

In fact, the method was deeply flawed: It de-emphasized the quality and content of the test items and focused instead on the rates of correct response obtained for previous test takers. A flurry of technical analyses showed that using the Golden Rule approach tends to produce tests with low reliability and to eliminate those test questions that best distinguish more proficient test takers from less proficient ones, reducing the validity of the assessment. In addition, the procedure is unlikely to eliminate items that are, in fact, biased and may distort the content of the test (Hackett et al., 1987; Linn & Drasgow, 1987; Marco et al., 1988). Yet another drawback is the fact that, in many cases, eliminating the more difficult test questions excludes those with the smallest group differences, a result contrary to the goals of the settlement (Linn & Drasgow, 1987). In the measurement community, an overriding concern was that the Golden Rule approach implicitly endorsed the idea that performance differences are in themselves evidence of bias (see the section “Fairness as Equality of Performance”).

ETS president Greg Anrig, who had agreed to the settlement, ultimately acknowledged that doing so was a mistake, given the technical concerns and the unanticipated attempts to generalize the approach beyond the tests involved in the litigation (Anrig, 1987a, 1987b). He also presented data seeming to show that the approach did not have the desired effect of increasing the passing rate for Black candidates on the insurance tests (Anrig, 1987b).

A positive effect of the Golden Rule case and its aftermath was the development and implementation of procedures for the detection of item bias (later to be renamed differential item functioning, or DIF) that were easy to use and understand (see Camilli & Shepard, 1994; Holland & Wainer, 1993). These methods, unlike the Golden Rule procedure, involve comparisons of the item performance of groups of test takers who have been matched in terms of the overall proficiency that the test is intended to measure (see the section “Differential Item Functioning”). Although bias-detection procedures had been developed earlier (e.g., Angoff & Sharon, 1974; Lord, 1977), the Golden Rule settlement and its subsequent repudiation spurred the widespread adoption of DIF detection procedures as a routine quality check.

Association of American Medical Colleges (AAMC) v. Carey (1990)

In 1979, New York State passed a law called the Standardized Testing Act, sometimes called the truth-in-testing law or the LaValle law, after the state senator who introduced it. The law required that testing agencies “file copies of test questions and answers with the Commissioner of Education’s office and that students be provided the opportunity to request copies of test questions, correct answer keys, scoring instructions, and their answer sheets” (McAllister, 1991, p. 20). The New York law followed similar legislation passed in California in 1978.

In *AAMC v. Carey*, the AAMC, sponsor of the Medical College Admission Test, sued New York State (whose governor was Hugh Carey) in U.S. District Court in 1980, setting off a chain of legal actions that were not resolved until 1999. The AAMC claimed that the Standardized Testing Act would compel it to quickly deplete the supply of usable Medical College Admission Test questions and that it was in conflict with federal copyright law. The AAMC was granted a preliminary injunction in 1980, exempting it from the law’s provisions. Other testing agencies, including ACT, College Board, ETS, the Graduate Management Admission Council, the Graduate Record Examinations Board, and the Law School Admission Council, initiated voluntary disclosure policies (McAllister, 1991) following negotiations about the number of test forms that needed to be released. In 1990, a federal district judge ruled that the New York law was indeed in direct conflict with federal copyright law and that the state could not compel the AAMC to disclose its test forms. At that point, test sponsors who had been voluntarily complying with the state filed suit, seeking to expand the court ruling and thus exempt themselves from the disclosure requirements (McAllister, 1991). However, the 1990 decision was subsequently overturned and all exemptions from it were withdrawn. After a lengthy period of litigation and negotiation between the state and the testing agencies, the court modified its original order to require that some but not all test forms be disclosed (Phillips, 2016). It was not until 1999 that all test disclosure lawsuits were dismissed (Zwick, 2002).

The suits challenging the truth-in-testing laws present an interesting example of a fairness issue that is not a question of equal treatment of test-taker groups. Instead, the case centers on a very general question of transparency: Do test takers have a right to examine at least some admissions test questions, along with their intended answers? Given the high-stakes nature of admissions tests and their role in distributing a valued asset—access to an educational program—it seems entirely reasonable to grant that right. It is now accepted practice for testing agencies to release a portion of admissions test questions, along with the answers.

Breimhorst v. Educational Testing Service (2000)

In 2000, Mark Breimhorst filed suit against ETS in federal court in California concerning an annotation that accompanied his GMAT score. (ETS was then the sponsor of the GMAT, which is now owned by a separate company.) Because Breimhorst had a disability—he had no hands—he took the test on a computer with a trackball instead of a mouse and was also granted 25% extra time. The notation that appeared with his scores said, “Scores obtained under special conditions” (Phillips, 2016).

The Breimhorst lawsuit sparked an industry-wide reconsideration of the policy of flagging test scores that had been obtained under nonstandard conditions. Flagging policies had long been a source of controversy. Proponents claimed that information about testing conditions was necessary for a correct interpretation of test scores and was also helpful in discouraging those who might request undeserved testing accommodations. Critics argued that flagging was a form of discrimination and that it stigmatized those with disabilities and violated their privacy (Zwick, 2002). The *Standards* acknowledges the difficulty of resolving this issue. According to the 2014 *Standards* (which largely echo the 1999 version in place at the time of the Breimhorst suit),

when there is clear evidence that scores from regular and altered tests or test administrations are not comparable, consideration should be given to informing score users, potentially by flagging the test results. . . . Where there is credible evidence that scores from regular and altered tests are comparable, then flagging generally is not appropriate. There is little agreement in the field on how to proceed when credible evidence on comparability does not exist. (AERA et al., 2014, pp. 61–62)

Despite these ambiguities, ETS agreed, as part of the 2001 settlement of the lawsuit, to discontinue flagging the scores of test takers who received extra time on the GMAT, GRE, TOEFL, and Praxis teacher certification tests. In 2002, after convening an expert panel on the subject, the College Board decided to stop flagging SAT scores obtained under nonstandard conditions and ACT stopped annotating scores obtained with extended-time administrations (Phillips, 2016).

Surprisingly, these flagging issues, which had apparently been resolved by 2002, re-emerged in 2018, when ACT, Inc., was sued by California test takers with disabilities, who claimed that ACT was providing information about their disability status and

test accommodations to colleges (Agrawal, 2020). According to an ACT spokesman, ACT had stopped annotating the test scores themselves with information about testing accommodations following the Breimhorst case, but had continued to relay information to colleges about test takers' disabilities if they had included that information in the survey that accompanies the test registration materials (Gewertz, 2018). The suit also claimed that ACT included test takers' disability status in a database sold to colleges to help them identify recruits. ACT agreed to stop providing all such information to schools and, in 2020, agreed to pay \$16 million to settle the case (Agrawal, 2020).

At the core of the Breimhorst case is the question of whether scores obtained under nonstandard conditions are comparable to those obtained in the usual way. Conducting rigorous research on this question is extremely difficult because there are typically very small numbers of test takers who share a particular disability and a particular type of nonstandard administration. In the Breimhorst case, the court placed the burden of proof on the testing agency to produce evidence that the test taker's scores were not comparable. By contrast, in a somewhat similar and roughly concurrent case, *Doe v. National Board of Medical Examiners* (1999), the court placed the burden on the test taker to show that his score, obtained under nonstandard conditions, was valid (Phillips, 2016). As Phillips noted, the 2014 *Standards* provides some clarity on this issue, stating that it is the responsibility of "test developers and/or users" to obtain and document evidence of the validity of score interpretations when a test is "changed to remove barriers to the accessibility of the construct being measured" (AERA et al., 2014, Standard 3.11, p. 68; see also Standard 9.9, p. 144).

Smith v. Regents of University of California (2019)

In December 2019, Kawika Smith, an African American high school student, along with three other students and six nonprofit organizations, sued the University of California [UC] and its then-president, Janet Napolitano, in California Superior Court, charging that the SAT and ACT, required for UC admission, "are demonstrably discriminatory against the State's least privileged students, the very students who would most benefit from higher education." (A separate case, for which the plaintiff was the Compton, California, school district, was consolidated with the Smith case.) The complaint goes on to say that the

mere presence of the discriminatory metric of SAT and ACT scores in the UC admissions process precludes admissions officers from according proper weight to meaningful criteria, such as academic achievement and personal qualities, and requires them instead to consider criteria that act as a proxy for wealth and race and thus concentrate privilege on UC campuses. (*Smith v. Regents of University of California*, 2019, p. 4)

The Smith lawsuit further claimed that admissions test scores contribute little to the prediction of college performance and that "affluent students can effectively purchase higher scores through expensive private tutoring services" (pp. 5–6).

In May 2020, the UC Regents voted to phase out existing admissions tests by 2025, contrary to the advice of the university's own Standardized Testing Task Force (University of California Academic Senate, 2020). The Regents' action was an endorsement of a recommendation by Janet Napolitano, who cited "the correlation of the SAT and the ACT to the socio-economic level of the student, and in some case[s], the ethnicity of the student" (Wilner, 2020). Test-optional policies, in which applicants decide whether to submit test scores, were to be put in place immediately, to be followed in 2023 by test-blind policies, in which test scores were not to be considered in admissions decisions even when submitted. However, a test-blind policy was initiated sooner than expected after the plaintiffs in the Smith case filed an additional complaint focusing on the inaccessibility of tests to applicants with disabilities because of pandemic-related disruptions. Under these circumstances, even test-optional policies were claimed to be unfair. The court ruled in favor of the plaintiffs and granted a preliminary injunction prohibiting the consideration of admissions test scores at any of UC's 10 campuses (Hoover, 2019; 2020), a decision that was subsequently upheld by an appeals court (Jaschik, 2020b).

A settlement of the Smith lawsuit was reached in May 2021, with UC agreeing to extend its test-blind policy through at least 2025 (*Settlement Agreement*, 2021). Later in the year, UC went a step further, announcing that the university would "continue to practice test-free admissions now and into the future" (Watanabe, 2021), requiring neither the SAT nor any other standardized test of its applicants. This statement appeared to foreclose the possibility, previously raised, that UC would develop a new admissions test.

The crux of the argument in *Smith v. Regents* was that

SAT and ACT scores are largely a proxy for a student's socioeconomic background and race. Instead of measuring students' academic ability or mastery of curriculum, SAT and ACT scores reflect demographic and socioeconomic characteristics like family income, parental education, and race. (p. 4)

The idea that admissions test scores are predictive of college grades only because both test scores and grades are influenced by SES has gained some traction (e.g., Geiser & Studley, 2002; Rothstein, 2004), but has been refuted by recent research. For example, Sackett and his colleagues conducted a series of meta-analyses in which they obtained partial correlations between test scores and subsequent grades, controlling for SES. They found that "across all studies, the admissions test retains over 90% of its predictive power when SES is controlled" (Sackett & Kuncel, 2018, p. 27).

Implicit in the argument advanced in the Smith lawsuit is the belief that a fair test will lead to equivalent performance by all relevant test-taker groups—in this case, groups defined by SES and race. (See the section "Fairness as Equality of Performance" for a discussion of this definition of fairness.) Admissions tests, which do not meet this criterion, are seen as an obstacle to the equitable consideration of college applicants. This perspective has been more widely embraced in the United States as a result of an

increased focus on racism and societal inequities, spurred by recent events. In addition, the coronavirus pandemic has magnified issues of test accessibility because of the closure of many test centers. The confluence of these factors has led to a powerful movement against the use of standardized admissions tests.

Conclusion: Legal Perspectives on Test Fairness

As Rebell (1989) noted more than 30 years ago, it is difficult to neatly summarize the impact of the courts on testing practices. In some cases, judicial intervention “has had the effect of upgrading the general level of psychometric practice throughout the industry,” while in others, judges have “venture[d] onto areas beyond their technical competence, with attendant detrimental results” (Rebell, 1989, p. 161). Rebell stressed the need to promote effective dialogue between the legal and psychometric professions as a way to maximize the chances of positive outcomes. As Phillips (2016) noted, better communication may also increase the chances of resolving issues outside the courtroom.

STATISTICAL PERSPECTIVES ON TEST FAIRNESS: MEASUREMENT, PREDICTION, AND LINKING INVARIANCE

During the civil rights era of the 1960s, the interest in using statistical analyses to aid in investigations of test fairness increased (Cole, 1993). In the early 21st century, statistical checks intended to examine test fairness are routine. In addition to general performance comparisons across demographic groups, these statistical analyses often include investigations of differential validity (e.g., Young, 2004) and DIF (Camilli & Shepard, 1994; Holland & Wainer, 1993). In selection contexts, fairness analyses typically focus on differential prediction of a criterion variable, such as performance in college (Young, 2004; Zwick, 2017a) or in an employment setting. In some testing situations, the factorial structure of tests and, more recently, the functions used to link one test form to another have been compared across groups as well.

Determining which groups should be explicitly considered in analyses of invariance or selection is complex. Historically, analyses of test fairness have focused on women and on racial and ethnic groups that are considered “protected” in the employment setting,⁶ but there is some disagreement about the rationale for focusing on these groups. In the educational research context, membership in particular racial or ethnic groups is sometimes viewed as a proxy for socioeconomic disadvantage (see Dorans & Cook, 2016). Novick and Ellis (1977) argued that, rather than accepting group membership as a surrogate for disadvantage, education and employment selection policies should consider “the explicit identification of those [candidate] attributes that constitute disadvantage” (p. 318). Policies that compensate for these disadvantages can then be crafted, they proposed. However, this viewpoint fails to account for the fact that members of some racial and ethnic groups, as well as women, people with disabilities, and members of various other groups, including those defined by age, national origin, language, or religion, may be subject to discrimination in selection

because of their group membership per se. Special consideration of members of these groups may therefore be appropriate even if they have not experienced disadvantage of the sort considered by Novick and Ellis, such as “deprivation of exposure to ‘educated’ persons, economic deprivation, or unstable home environment” (p. 319). The difficulty in determining which groups require special scrutiny in educational testing and selection is evidence of the degree to which values play a role in evaluations of validity and fairness.

In general, the goal of analyses of measurement invariance is to verify that the test “works the same way” for all groups. Statistical analysis results can never be sufficient to demonstrate fairness; other aspects of the testing situation, including test content and administration conditions, and in some cases, opportunity to learn, must be considered as well. Analysis results, however, can serve to bolster the case for a test’s fairness or to provide evidence that the test may not, in fact, be fair.

Some fairness analyses—in particular, DIF and factorial invariance analyses—are best understood from the perspective of measurement invariance. According to Millsap (2011, p. 1), “measurement invariance is built on the notion that a measuring device should function in the same way across varied conditions, so long as those varied conditions are irrelevant to the attribute being measured.” Millsap offers an example from the realm of physical measurement: A weight scale should not yield different results for two objects that have the same weight, but vary in shape—an irrelevant characteristic. A departure from measurement invariance is of particular concern when the irrelevant factor is subpopulation membership. This is reflected in the definition offered by Liu and Dorans (2016), who noted that measurement invariance “refers to the degree to which an instrument measures an entity in the same way across different subpopulations” (p. 84).

Although it may seem to be an aspect of measurement invariance, prediction invariance (i.e., the absence of differential prediction) is best considered a separate phenomenon. As detailed in the “Prediction Invariance” section, a “deep inconsistency” between measurement and predictive invariance can be shown to exist in many plausible situations (Millsap, 2011, p. 299). In the following sections, measurement invariance (including DIF and factorial invariance) is considered first, followed by prediction invariance and linking invariance. Invariance of validity coefficients is discussed in the context of prediction invariance. Examples from the research literature are included to illustrate the kinds of information that these analyses can yield.

Measurement Invariance

The most stringent definition of measurement invariance requires that, for two population groups, G_1 and G_2 ,

$$P(\mathbf{X}|\boldsymbol{\theta}, G_1) = P(\mathbf{X}|\boldsymbol{\theta}, G_2), \quad (1)$$

where $\mathbf{X}$ is a p -dimensional vector random variable consisting of the scores on p observable measures such as test or item scores; $\boldsymbol{\theta}$ is an r -dimensional vector random variable, $1 \leq r \leq p$, representing the unobservable characteristics that the measures are

intended to assess; and $P(A|B, C)$ represents the conditional distribution of A , given B and C . Equation 1 states that the conditional distribution of the observed measurements, given the (possibly multidimensional) target construct, does not vary across groups. As an example, X could represent item scores on the math section of the SAT, θ could be a set of unobservable true math proficiencies (e.g., algebra, problem-solving, and geometry), and G_1 and G_2 could represent native English speakers and nonnative English speakers. If Equation 1 held, that would mean that the distribution of math item scores, given the test takers' true math proficiencies, was the same for these two groups. For simplicity, only the two-group case is described in the sections that follow.

An important assumption in evaluating the plausibility of the hypothesis of Equation 1 is that the conditions of measurement, such as the timing, instructions, and administration mode, are the same for both groups (Dorans & Middleton, 2012). Unless these conditions are the same, any departures from invariance cannot be attributed to group membership per se.

Differential Item Functioning

In a typical DIF analysis, both X and θ in Equation 1 are univariate. X represents an item score and θ is the construct the test is intended to measure, such as math proficiency or knowledge of U.S. history. A separate DIF analysis is conducted for each item.⁷ Only the case of dichotomously scored items is discussed here. DIF analysis methods for polytomous items are discussed by Penfield and Camilli (2007) and Zwirk et al. (1997). For dichotomous items, the purpose of a DIF analysis is to determine whether the probability of a correct response is the same for individuals from two groups who are matched with respect to θ . Typically, only the items on the test itself are available as measures of θ . In essence, then, the item being assessed for DIF—the *studied item*—is scrutinized to determine whether it measures something different from the construct that is assessed by the entire set of test items, considered as a whole. This aspect of DIF analysis—the fact that each item is evaluated with respect to the test as a whole—has been pointed out as a weakness of DIF analysis (e.g., Camilli & Penfield, 2007), and it is indeed true that DIF procedures would not identify any items as problematic if they were all similarly biased. Instead, DIF analyses are useful in identifying items that assess a skill or characteristic that is not the intended target of measurement.

Suppose that some individuals taking an algebra test are English learners, while others are native speakers of English. Most of the test questions contain only equations, but one includes a substantial amount of English text. Assuming that an adequately powerful statistical procedure were used, this item would be found to show DIF that was unfavorable to the English learners under the following conditions: (a) responding correctly to the item required a skill, comprehension of English (say, θ'), that was outside the intended construct, algebra proficiency (θ); (b) this auxiliary skill was not measured by the remaining items on the test; and (c) the conditional distribution of this secondary skill, given the intended construct, differed across the two language groups. That is,

$$P(\theta' | \theta, G_1) \neq P(\theta' | \theta, G_2). \quad (2)$$

(see Millsap 2011, p. 55). Dozens of DIF analysis methods have been developed since the 1970s. (Earlier methods were referred to as *item bias* detection methods. Researchers gradually switched to the more neutral term, *differential item functioning*, to reflect the fact that test-taker groups could differ in their conditional item performance for many benign reasons, such as differences in interests or in course background.) A key dimension on which these DIF methods vary is the manner in which the construct of interest is measured for purposes of matching test takers. Some methods are based on particular item response theory (IRT) models and involve attempts to compare the item performance of two groups, conditional on θ , the unobserved construct assessed by the test. Others match test takers on observed test scores and do not require any assumptions about the form of the item response function.⁸ The main categories of DIF methods currently in use are briefly outlined here and some key issues in DIF analysis are considered. The required computations are described by Camilli (2006), Millsap (2011), and Penfield and Camilli (2007).

In the case of dichotomous item i with score X_i , the hypothesis of interest in DIF analysis takes the form of Equation 1. More specifically, the hypothesis is

$$\Pr(X_i = 1 | \theta, G_1) = \Pr(X_i = 1 | \theta, G_2) \text{ for all } \theta, \quad (3)$$

where $X_i = 1$ indicates a correct score and $X_i = 0$ indicates an incorrect score. Equation 3 states that the probability of correct response is the same for the two groups at all θ levels; that is, the item response functions for Groups 1 and 2 are identical. This is Lord's (1980) definition of the absence of item bias. In terms of the earlier example, the hypothesis in Equation 3 would be satisfied if, at every level of algebra proficiency (θ), the native and nonnative English speakers had the same probability of correct response on the item.

One way to test the hypothesis in Equation 3 is to use likelihood ratio tests to compare the fit of competing IRT models (Thissen et al., 1993). A specific IRT model (ordinarily a logistic or normal ogive model) is assumed to hold. All items except the studied item are assumed to be free of DIF and therefore to have item parameters that are common across the two groups. Then the fit of a model that constrains the item parameters for the studied item to be the same for the two groups is compared to the fit of a model that does not include those constraints. If the less-constrained model fits better to some pre-determined degree, the item is assumed to have DIF. Other IRT-based methods include graphical comparison of the item response functions of the two groups and evaluations of the area between the item response functions.

The most widely used observed-score approaches for DIF detection are the Mantel-Haenszel (MH; Holland & Thayer, 1988; Mantel & Haenszel, 1959), standardization (Dorans & Kulick, 1986), and logistic regression (Swaminathan & Rogers, 1990) methods.⁹ The MH and standardization statistics are computed based on a $2 \text{ (groups)} \times 2 \text{ (item scores)} \times K$ contingency table, where K is the number of levels of the total

test score, $S = \sum X_i$. The importance of including the studied item when computing S is discussed by Holland and Thayer (1988) and Zwick (1990). It is important that the matching variable be free of any DIF items (other than possibly the studied item). Therefore, in MH and standardization DIF analyses, it is common practice to use a refinement procedure (see Zwick, 2012). This involves conducting a preliminary analysis to identify any DIF items on a test. These items are then removed from the matching variable, with the exception that the score on the studied item must always be included. When refinement is used, S is redefined as the sum of the scores on all DIF-free items on the test, plus the score on the studied item.

In both the standardization and the MH procedures, the difference between the studied item scores of the two groups, conditional on S , can be tested for statistical significance. The standardization-based DIF statistics, however, are often used purely for descriptive purposes (Zwick, 2012). The most frequently used effect size measure associated with the standardization approach is the difference between the observed proportion of correct responses on the studied item for the group of primary interest (say, Group 1) and the predicted proportion of correct responses that would have been attained by Group 2 if its distribution on S had been the same as that of Group 1 (Dorans & Kulick, 1986). The MH method yields an effect size measure that is based on the ratio of the odds of getting the item right for Group 1 to the odds for Group 2, conditional on S ; it is an assumption of the MH approach that this odds ratio is constant over all levels of S . A system for classifying DIF severity into three levels, labeled A (negligible), B (slight to moderate), and C (moderate to large), was developed at ETS and is now widely used. The criteria for classifying items are based on statistical significance tests in combination with effect size measures (see Zwick, 2012, pp. 3–5).

The logistic regression approach to DIF detection involves fitting a model to predict the item score, X_i , in which the predictors are S , G (group membership), and the $G \times S$ interaction. If the interaction exceeds a criterion value, indicating that the predictive value of S varies across groups, the item is judged to have DIF. The logistic regression approach can yield effect size measures as well (Penfield & Camilli, 2007).

Observed-score DIF methods like the MH, standardization, and logistic regression test a hypothesis that can be viewed as an approximation to the hypothesis in Equation 3. Specifically, they test the hypothesis that two groups are equally likely to answer an item correctly, conditional on test score S :

$$\Pr(X_i=1|S, G_1) = \Pr(X_i=1|S, G_2) \text{ for all } S. \quad (4)$$

Except in special circumstances, the hypotheses in Equations 3 and 4 are not equivalent. That is, test takers who are matched in terms of total score S will not, in general, be matched on θ . Suppose that Group 2 has a lower mean on θ than Group 1, that is, $E(\theta|G_2) < E(\theta|G_1)$, and that test reliability is the same for the two groups. Then, under the assumptions of classical test theory, $E(\theta|S, G_2) < E(\theta|S, G_1)$. Only when S is a sufficient statistic for θ does this matching problem disappear.¹⁰ Sufficiency occurs for the Rasch model (and the partial credit model, in the case of polytomous items),

but not for other common IRT models (Holland & Thayer, 1988; Zwick, 1990; Zwick et al., 1997). The higher the reliability of S , the smaller the practical effect of the difference between the hypotheses of Equations 3 and 4.

Another well-established DIF detection method, the SIBTEST procedure (Shealy & Stout, 1993), is neither an IRT-based method nor, strictly speaking, an observed-score method. It can be applied to a single item or to a set or “bundle” of items. When applied to one item, it is similar to the standardization approach except that it includes an adjustment to the matching variable that is intended to avoid errors in matching caused by measurement error in the observed score.

A DIF ANALYSIS EXAMPLE In an early application of the MH DIF method, Zwick and Ercikan (1989) analyzed U.S. history items that were administered to a nationally representative sample of 11th graders as part of the National Assessment of Educational Progress (NAEP) in 1986. More than 7,000 students took one of four history blocks, each consisting of 34 to 36 items. In the initial analyses, test takers were matched on their number-right scores on the history blocks to which they responded; subsequent analyses also took into account students’ course background in history. Here, we consider only the items that showed moderate to large DIF—C items in the ETS classification—in the initial analyses that compared Hispanic and Black test takers, respectively, to White test takers. In the Hispanic–White comparison, the only C item, which was differentially easier for Hispanic test takers, was about Latin American and Asian immigration to the United States. The Black–White comparison yielded three C items favoring Black test takers; the topics were Martin Luther King, Harriet Tubman, and the Underground Railroad. In neither of these analyses was there any C item that was differentially easier for White test takers. These DIF findings lend themselves to at least two possible explanations. One is that White students were less likely to have attended history classes that included the topics in question or to have learned the material outside school; the other is that these topics were of less interest to White students than to Hispanic and Black students with the same number-right score, leading to inferior performance.

These results illustrate why a policy of simply deleting all items that show DIF would be unwise. The DIF items in the U.S. history assessment addressed topics that were clearly part of the intended construct; the DIF was probably caused by differences in exposure to the material or in prior knowledge or interests. In typical assessment programs, items found to have DIF are reviewed by subject area experts from diverse backgrounds to determine whether item responses are likely to have been affected by some construct-irrelevant factor. If the review takes place during a pilot test phase, items can be revised or simply eliminated.

In most instances, the reasons for DIF findings are far less apparent than in the NAEP example. However, in studies conducted in the late 1980s and early 1990s, certain patterns emerged. For example, DIF favoring men was found on items about scientific topics, sports, and military activities, while DIF favoring women was found on questions about human relationships and the arts (see Zwick, 2002). The associations between

test content and DIF that were observed in these early studies had an impact on test development practices (Zieky, 1993) and on the sensitivity review procedures that are now routine practice at testing companies (see Ramsey, 1993). Whether these patterns would hold today is an open question. Little research of this kind has been conducted in recent years.

Although the vast majority of DIF research has been observational in nature, Schmitt et al. (1993) pointed out that more useful information can be obtained from randomized DIF studies that involve items specially constructed to test particular DIF hypotheses. In one such study, for example, Schmitt et al. (1988) found convincing evidence that *true cognates*—in this case, English words that were similar to Spanish words with the same meaning—facilitated the performance of Hispanic test takers. In particular, “the true cognates that consistently made the items differentially easier for Hispanics were words with a higher frequency of usage in the Spanish language” (Schmitt et al., 1993, p. 301).

Factorial Invariance

In educational testing, analyses of factorial invariance are far less common than DIF analyses, which are now considered standard quality-control procedures. Analyses of factorial invariance are more likely to be performed in the fields of intelligence and personality testing or in other areas in which a particular factor structure has emerged from psychological theory.

In analyses of factorial invariance, both $\mathbf{X}$ and $\boldsymbol{\theta}$ in Equation 1 are multivariate and continuous. $\mathbf{X}$ is a $p \times 1$ vector of item or subtest scores, and $\boldsymbol{\theta}$ is an $r \times 1$ vector of scores on the latent common factors assumed to underlie these scores. Here, we assume $r < p$ (i.e., a strict inequality holds). In the simplest case, $\mathbf{X}$ is assumed to be related to $\boldsymbol{\theta}$ through a linear factor model within each group:

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\tau}_G + \boldsymbol{\Lambda}_G \boldsymbol{\theta} + \mathbf{U}, \\ E(\mathbf{X}|\boldsymbol{\theta}, G) &= \boldsymbol{\tau}_G + \boldsymbol{\Lambda}_G \boldsymbol{\theta}, \\ \text{Cov}(\mathbf{X}|\boldsymbol{\theta}, G) &= \text{Cov}(\mathbf{U}|G),\end{aligned}\tag{5}$$

where $\boldsymbol{\tau}_G$ is a $p \times 1$ vector of latent factor intercepts for group G , $\boldsymbol{\Lambda}_G$ is the $p \times r$ factor pattern matrix for group G , and $\mathbf{U}$ is a $p \times 1$ vector of scores on the p unique factors. The unique factors are assumed to be uncorrelated with the common factors $\boldsymbol{\theta}$ and with each other. $\text{Cov}(\mathbf{U}|G)$ is a diagonal matrix whose elements are the p variances of the unique factors in group G . The factor model in Equation 5 says, in essence, that the observed test scores can be expressed as a weighted sum of the underlying factors, plus an intercept term and a residual term. For example, $\mathbf{X}$ might contain scores on a set of math tests and the factors $\boldsymbol{\theta}$ might correspond to broad areas of mathematical proficiency such as algebra, geometry, trigonometry, and calculus.

Under the assumption that $\mathbf{X}|\boldsymbol{\theta}$ is multivariate normal, the following two conditions are sufficient to establish measurement invariance in the two-group case (Millsap, 2011):

$$\begin{aligned} E(\mathbf{X}|\boldsymbol{\theta}, G_1) &= E(\mathbf{X}|\boldsymbol{\theta}, G_2) \\ \text{Cov}(\mathbf{X}|\boldsymbol{\theta}, G_1) &= \text{Cov}(\mathbf{X}|\boldsymbol{\theta}, G_2) \end{aligned} \quad (6)$$

It is not necessary to assume that the common factor means and covariance matrices are invariant across populations; the invariance property applies to the conditional distribution of the observed variables $\mathbf{X}$, given $\boldsymbol{\theta}$.

If multivariate normality does not hold, additional conditions may be needed to establish the most stringent type of factorial invariance because higher order moments of the conditional distributions of $\mathbf{X}$ given $\boldsymbol{\theta}$ may differ across groups even when the conditions of Equation 6 hold. Weaker forms of factorial invariance, which do not require identity of the conditional distributions, are described by Millsap (2011) and Liu and Dorans (2016). Statistical techniques used to investigate factorial invariance include exploratory and confirmatory factor analysis and structural equation modeling.

A FACTORIAL INVARIANCE STUDY EXAMPLE An example of a factorial invariance analysis is a study by Chiorri et al. (2016), who sought to investigate the factor structure of the Big Five personality factors (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience), as well as its invariance across gender. According to the authors, these factors are so well established that they constitute “a universal language in personality psychology” (Chiorri et al., 2016, p. 88). The authors administered a 44-item measure called the Big Five Inventory to a large Italian community sample and then analyzed the results using exploratory structural equation modeling, described as an integration of confirmatory and exploratory factor analysis. They found the expected five-factor structure (although with some minor departures from simple structure, in which each item loads on only a single factor) and determined that the gender comparisons met the invariance criteria of Equation 6. Although measurement invariance does not require that the factor means and covariance matrices be the same across groups, the authors tested these conditions as well. They concluded that the factor covariance matrices, but not the factor means, were the same for men and women. Women scored higher on all factors except Openness to Experience. The authors argued that their methodological approach was an improvement over more traditional confirmatory factor analysis methods and that their study therefore provided more appropriate tests of factorial invariance as well as more reliable evidence of gender differences in the personality factors measured by the Big Five Inventory.

Prediction Invariance

In many contexts, including admissions testing and employment testing, scores are used to predict a criterion measure using linear regression models that take the following form for Group G :

$$\begin{aligned} Y &= \alpha_G + \boldsymbol{\beta}_G' \mathbf{X} + U, \\ E(U) &= 0, \text{ and } \text{Var}(U) = \sigma_G^2, \end{aligned} \quad (7)$$

where Y is a criterion measure (assumed here to be unbiased), α_G is the intercept term, β'_G is a $1 \times p$ vector of coefficients, $\mathbf{X}$ is a $p \times 1$ vector of predictors, U is an error term, and p is the number of predictors. For example, in a typical study of college admissions testing, the predictors $\mathbf{X}$ are admissions test score and high school GPA, the β_G are the weights for combining the predictors, and the criterion Y is first-year college GPA.

Under the linear regression model (which is the only prediction model we consider here), the predicted value of Y , given $\mathbf{X}$, is

$$\hat{Y} = E(Y|\mathbf{X}) = \alpha_G + \beta'_G \mathbf{X} \quad (8)$$

Under prediction invariance,

$$P(Y|\mathbf{X}, G_1) = P(Y|\mathbf{X}, G_2). \quad (9)$$

Equation 9 says that the distribution of Y , given $\mathbf{X}$, is the same across groups. If the regressions were not the same across groups in the college admissions testing example, then the predicted college GPA would, in general, differ for members of Groups 1 and 2 who had the same test score and high school GPA.

Under normality of $Y|\mathbf{X}$, satisfying the invariance condition in Equation 9 requires only that

$$(\alpha_1, \beta'_1, \sigma_1^2) = (\alpha_2, \beta'_2, \sigma_2^2), \quad (10)$$

which means that the intercepts, regression coefficients, and residual variances are the same for Groups 1 and 2. Although some researchers have discussed the importance of equality of the residual variances σ_G^2 (see Einhorn & Bass, 1971; Jensen, 1980), particularly under certain selection rules, most studies of differential prediction focus on the less stringent requirement that

$$(\alpha_1, \beta'_1) = (\alpha_2, \beta'_2), \quad (11)$$

which Millsap (2011) called regression invariance.

Regression invariance is at the root of one of the most commonly invoked definitions of bias in the field of educational measurement and statistics. Drawing on a principle suggested by Humphreys (1952), Cleary (1968) stated in a now-classic article that

[a] test is biased for members of a subgroup . . . if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. . . . The test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance. (p. 115)

The *Standards* includes a similar statement but departs from Cleary's viewpoint in that it does not identify differential prediction as evidence of test bias:

The term *predictive bias* may be used when evidence is found that differences exist in the patterns of associations between test scores and other variables for different groups, bringing with it concerns about bias in the inferences drawn from the use of test scores. (AERA et al., 2014, p. 51)

This distinction reflects the fact that factors other than test bias may produce departures from regression invariance. It is well known, for example, that measurement error in the predictors can lead to violations of prediction invariance (e.g., see Linn & Werts, 1971). Suppose a test score X , which is a measure of θ , is used to predict Y for two groups of test takers. If X is not perfectly reliable and if the groups have different means on θ , then members of the two groups who are matched on X are not, in fact, matched on θ . These errors in matching, analogous to those described in the section on DIF, will generally cause the observed-score regressions to differ even when the regression of Y on θ is the same for all groups. In fact, measurement error in the predictors can fully explain the overprediction of the criterion variable for lower scoring groups, a phenomenon that has been replicated consistently in the testing literature since the 1970s. In particular, it is frequently found that when admissions test scores and high school GPA are used to predict college GPA, the use of a single regression model will result in predicted college GPAs that, on average, exceed the actual college GPAs for Black and Latino students, a result that many outside the measurement community find surprising. (Factors other than measurement error, such as quality of previous schooling, may contribute to this recurrent pattern as well; see Zwick, 2002, 2017a; Zwick & Himelfarb, 2011).

Model underspecification can also lead to differential prediction. The admissions literature again serves as a useful example. The prediction models applied in most admissions research typically include only admissions test scores and previous grades as predictors because these are the factors most often used by postsecondary institutions. The existence of other important predictors of college GPA that have distributions that vary across groups (conditional on the included predictors) will produce violations of regression invariance (Linn, 1973; Linn & Werts, 1971) in these sparse models. In addition, regression models that are identical in the population can be rendered unequal via selection, and different regressions can be rendered equal (Linn, 1984). Finally, the criterion variable Y could itself be biased (Linn, 1973), causing the regressions to differ, although we have assumed that is not the case in this section.

How does prediction invariance differ from measurement invariance? To illustrate the inconsistency between these two phenomena, Millsap (2011, pp. 285–300) reformulated the regression parameters (Equation 10) in terms of the parameters of factor models in which X and Y are considered jointly (i.e., in the model of Equation 5, Y is treated as though it is an additional X variable). He began with the case in which a single factor underlies (X, Y) . Among his conclusions are the following: When factorial invariance (Equation 6) holds and there are no group differences in common factor variances, slope invariance will hold but intercept invariance will not: $\beta'_1 = \beta'_2$, but $\alpha_1 \neq \alpha_2$. The predictive bias in this situation will penalize the higher scoring group if a common regression equation is used for prediction. Conversely, it is possible

for regression invariance to hold when measurement invariance is violated. Millsap showed that these types of inconsistencies are not anomalies, but can occur under very general conditions. That is, “the presence or absence of group differences in regression intercepts is not diagnostic for the presence or absence of measurement bias” (Millsap, 2011, p. 291). He recommended that latent variable models be used to explore the causes of both measurement and prediction invariance (see, e.g., Culpepper et al., 2019).

EXAMPLE OF A PREDICTION INVARIANCE STUDY In a study of about 8,000 high school sophomores who participated in the High School and Beyond survey, Zwick and Sklar (2005) investigated the degree to which prediction of first-year college GPA varied across four student groups that were based on ethnicity and first language (Black/English, Hispanic/English, Hispanic/Spanish, and White/English). The strategy followed, which was based solely on observed scores, is sometimes called moderated multiple regression (e.g., Millsap, 2011). It involves fitting a series of nested regression models. The first model, Model 1, included only the two academic predictors, high school GPA and total SAT score. Model 2 added indicators for group membership, and Model 3 added all two-way interactions between group membership and the academic predictors.

Model 1 provides a test of whether the academic predictors contribute to the prediction of college GPA (Zwick & Sklar, 2005). Assuming Model 3 does not subsequently produce a statistically significant increase in the value of R^2 (which measures the degree of association between GPA and the predictors), Model 2 provides a test of the hypothesis of equal intercepts across groups, the hypothesis of greatest interest in analyses of this kind. Model 3 provides a test of whether the strength of the relationship between college GPA and the two academic predictors varies across groups—a test of equal slopes. (If there is evidence of different slopes, the test of equal intercepts is not meaningful.)

The Model 1 results showed that both academic predictors were statistically significant, jointly producing an R^2 value of .22, which indicates that 22% of the variability in college GPA was accounted for by these predictors (Zwick & Sklar, 2005). The R^2 obtained for Model 2 showed a statistically significant increase from the value obtained for Model 1, but the Model 3 R^2 was not statistically different from the value obtained for Model 2. Therefore, the Model 2 results could legitimately be interpreted as a rejection of the hypothesis of equal intercepts. In particular, the study found that the Black/English group had a lower intercept than the White/English group, which served as a reference group. This means that a member of the Black/English group would have a lower predicted college GPA than a member of the White/English group with the same high school GPA and SAT score. A regression equation based on both groups combined would overpredict the college GPAs of the Black/English group. There was no statistically significant evidence of intercept differences for the two Hispanic groups.

DIFFERENTIAL VALIDITY Differential validity analyses involve the comparison of validity coefficients (test-criterion correlations) across groups (e.g., Young, 2004). These analyses will not necessarily yield the same conclusions about invariance as differential prediction analyses because identity of regressions does not in general imply identity of correlations, nor does identity of correlations imply identity of regressions.¹¹ Because of its implications for selection in some admissions and employment testing situations, prediction invariance is arguably more relevant to fairness than is invariance of validity coefficients. Nevertheless, it is generally of interest to assess the magnitude of the test-criterion correlations across key demographic groups, and this information is typically included in technical documentation for tests. The *Standards* indicates that test developers “should include relevant subgroups in validity, reliability, precision” and other studies and that “if sample sizes permit, it is often valuable to carry out separate analyses for relevant subgroups of the population” (AERA et al., 2014, p. 64).

Linking Invariance

On tests for which multiple forms exist, linking procedures must be used to ensure that scores on the alternate versions can be used interchangeably. Linkages also come into play when scores on one test may be substituted for scores on another. For example, in the case of the ACT and SAT, both used in college admissions, university officials are interested in knowing how scores on the two tests are related. When linkages must be established among different tests or test forms, the question of whether linking functions are invariant across groups becomes relevant.

Linking invariance differs from both measurement invariance and prediction invariance. Borrowing the notation used by Liu and Dorans (2016), linking invariance holds if

$$L(Y \leftarrow X|X, G_1) = L(Y \leftarrow X|X, G_2), \quad (12)$$

where X and Y represent the scores on two tests or two versions of a single test. Equation 12 says that the function used to link X to Y is the same for Groups 1 and 2.

Kolen (2004) reviewed the literature on population invariance in equating and linking, which dates back to the 1950s. (Equating is a particular kind of linking in which the various versions of a test have been developed using the same content and statistical specifications; see Dorans & Holland, 2000, for further detail.) His conclusion was that equating theory, “in combination with empirical research suggests . . . that when alternate test forms are constructed so that they are very similar in content, difficulty, and reliability, population invariance holds approximately” (Kolen, 2004, p. 11). However, invariance cannot be assumed to hold; it must be investigated for each linkage and test-taker group.

Dorans (2004) recommended that linking invariance be assessed as a routine component of fairness analysis for tests that exist in alternate forms. Dorans and Holland (2000) proposed two indexes that could be used to assess the degree to which a set of linking results departed from invariance. One is a root-mean-square

difference (*RMSD*) measure that is computed for each value of Y . The *RMSD* is computed by first obtaining the squared difference between the results of applying each subpopulation linking function and the overall population linking function at $Y = y$. The square root of the weighted sum (over subpopulations) of these squared differences is obtained and then divided by the population standard deviation of X . The second index, called the root expected-mean-square difference (*REMSD*), is obtained by averaging the *RMSD* values over the distribution of Y before taking the square root. This yields an overall summary measure of the departure from linking invariance. Additional indexes are discussed by Liu and Dorans (2013). In general, the linking functions for particular groups are compared to an overall (population) linking function. Departures from invariance are then evaluated to determine whether they are practically meaningful.

EXAMPLE OF A LINKING INVARIANCE STUDY As an illustration of linking invariance analysis, Dorans and Holland (2000) studied the invariance of several ACT–SAT linkages over gender groups. (In this analysis, the linking functions for men and women were compared; a population linking function was not obtained.) Under the parallel-linear linkage models used in this analysis, the *RMSD* does not vary over values of Y and, therefore, the *REMSD* and *RMSD* statistics are equivalent. Of particular interest is the linkage between the ACT Composite score and the sum of the SAT Verbal and Math scores ($SAT\ V + M$), which had a correlation of .916. For assessing invariance over gender groups, the *RMSD* was found to be 5.4% of a standard deviation unit. This translated to 0.3 points on the ACT Composite scale and 10.5 points on the $SAT\ V + M$ scale, which Dorans and Holland judged to be small deviations. For the linkage between the math sections of the ACT and SAT, which had a correlation of .892, the *RMSD* was found to be even smaller, at 1.5% of a standard deviation unit. By contrast, the *RMSD* for the linkage between the ACT English and SAT Verbal sections, which had a correlation of .83, was much larger, at 7.4%. Dorans and Holland stressed that these *RMSD* values pertained only to invariance over gender groups and that there could be “other subgroups that will exhibit even larger linear linking function differences than those for males and females” (p. 296). It is important to note that linkages between different tests (e.g., the ACT and the SAT) are likely to result in larger departures from linking invariance than linkages among forms of a single test.

SELECTION FAIRNESS

Fairness in selection, which is relevant mainly in admissions and employment contexts, is a complex topic because selection decisions typically involve not only test scores, but also additional characteristics or accomplishments of the candidates. Decisions may also depend on policies imposed by governments or institutions. In addition, selection processes may involve predictions of future performance.

Regression models such as those discussed in the section on “Prediction Invariance” often play a role in the process of selecting candidates for employment and for admission to higher education. As noted in that section, determining whether regressions are invariant across groups is not a straightforward process because of the impact of measurement error, omitted predictors, and selection effects. In the decade following the publication of Cleary’s (1968) landmark article, a host of “culture-fair” selection procedures were proposed as alternatives to the regression approach. It is notable that this area of study was motivated by an openly stated belief that the use of standardized tests was in some sense unfair. In the lead article of a special issue of *Journal of Educational Measurement* devoted to this topic, Petersen and Novick (1976, p. 3) noted that the goal of this area of study was to “eliminate cultural or racial unfairness arising from the use of tests” in selection.

Detailed reviews of the selection fairness research published during this time appear in Camilli (2006) and Petersen and Novick (1976); here, I briefly review some of the key contributions. It is assumed that two groups of test takers, one of which is at risk for being treated unfairly, are given a test that is intended to predict success on a criterion measure, such as job performance or grades earned in higher education. In the literature of the time, the groups were often labeled “minority” and “majority” and the minority group was assumed to have lower averages on the test and the criterion measure. Success on the criterion was typically conceptualized as performance exceeding a particular value. Each fairness method provided a way of determining what cut scores should be used on the test for selection purposes.

Einhorn and Bass (1971) recommended that each group’s cut score be chosen so that the risk of failure on the criterion at the cut score is the same across groups. Even if the two groups have identical regression slopes and intercepts, resulting in fair selection according to Cleary’s definition, selection is unfair according to Einhorn and Bass if the two groups have different residual variances. That is, the invariance condition in Equation 10, not merely the condition in Equation 11, is claimed to be necessary for fair selection.

Robert Thorndike (1971) demonstrated that a selection rule that is fair according to the Cleary model will generally be unfair to the lower scoring group “in the sense that the proportion qualified on the test will be smaller, relative to the higher-scoring group, than the proportion that will reach any specified level of criterion performance” (p. 63). He proposed a *constant ratio model*, which requires that cut scores for each group be chosen so that the ratio of the proportion selected to the proportion successful on the criterion is the same across groups. It is possible for a selection rule that is biased under the Cleary model to be fair according to the Thorndike model. If the distribution of the criterion measure differs across groups, the two models will agree that a selection rule is fair only if the two groups have the same regression slopes and intercepts and the test has perfect validity (Petersen & Novick, 1976).

Cole (1973) suggested that, in evaluating fairness, the focus should be on the conditional probability of selection, given success on the criterion. According to Cole,

[the] basic principle of the conditional probability selection model is that for both minority and majority groups whose members can achieve a satisfactory criterion score . . . there should be the same probability of selection regardless of group membership. (p. 240)

Some models proposed during the culture-fair selection era required the user to assign numerical values to particular outcomes based on their perceived desirability from a social justice perspective. For example, Darlington (1971) developed a framework involving three variables—a predictor, a criterion, and a group membership variable. He proposed a selection model that would require the user to specify, in quantitative terms, the value of choosing individuals from particular cultural groups. Darlington's model would, in effect, "add a specified number of points to the [predictor] scores of one subpopulation and then use the same predictor cut score" for all groups (Petersen & Novick, 1976, p. 23).

Another type of selection model that requires the user to assign values to outcomes based on social justice principles is the decision theory approach proposed by Gross and Su (1975) and Petersen (1976). Petersen assumed that minority and majority group members are to be considered for selection based on a test score, and "qualified" candidates are those who have the potential to be successful. Utility values are assigned to each possible selection decision (e.g., "accept a qualified candidate who is a minority group member"), with higher values representing more desirable outcomes. Principles of decision theory are then used to calculate the cut scores for each group that maximize the expected utility.

None of the fairness models put forward during the 1970s was widely applied outside the research community. The proposed selection procedures were controversial—as they would be today—because they embodied particular social justice goals and required that, in order to promote these goals, qualifying scores on selection measures be allowed to vary across demographic groups. Today, any such methods would be prohibited in the employment context because of a provision in the Civil Rights Act of 1991 that states that it is

an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of employment related tests on the basis of race, color, religion, sex or national origin. (Civil Rights Act, 1991, Title VII)

The selection methods proposed in the 1970s had a further drawback that compounded any possible political or philosophical objections: The techniques were complex and, in some cases, internally contradictory in the sense that reformulating them in terms of the proportions of individuals rejected, rather than the proportions selected, led to different selection rules (Petersen & Novick, 1976). As Cole stated in 1981, "the large amount of technical work on bias in the last 15 years cannot be said

to have significantly clarified the public controversies or to have resolved many issues associated with bias in the public mind” (p. 1067).

Despite the failure to produce a consensus about the meaning of fairness or to improve communications with the public on this topic, the focus on statistical approaches to fairness during this period did have a positive effect: It led to creative thinking about the ways in which statistical analyses can be used to support investigations of equity. Partly as a result of the research conducted during this period, examinations of differential prediction, differential validity, and DIF are now widely regarded as necessary steps in evaluating fairness in measurement and selection.

More recently, researchers have proposed additional ways of using test scores in higher education admissions that are intended to mitigate any possible negative effects on disadvantaged applicants, though most of these methods have been applied only in a research setting. An approach that attracted media attention in the late 1990s was the Strivers method, developed by an ETS vice president. It was described as a way to identify “students who score better than expected on SATs, given their economic and social backgrounds,” with the goal of awarding them “extra credit” (Carnevale & Haghighat, 1998, p. 123). A candidate’s SAT score was predicted using a regression analysis in which the predictors were socioeconomic factors, such as parents’ income and education and school quality. Candidates whose actual SAT scores exceeded their predicted scores by more than 100 points were considered strivers. After the idea was publicized, controversies quickly erupted, and the method was never implemented. Regression-based approaches in the same spirit had been proposed much earlier (Manning, 1977); others appeared later (Gaertner & Hart, 2013; Studley, 2004; Young & Johnson, 2004).

Indexes of disadvantage have also been proposed as ways of providing context for test scores (Bastedo et al., 2021; Blatter & Glynn, 2016; Gaertner & Hart, 2013; Goggin, 1999; Stricker et al., 2002; Zwick et al., 2021), and constrained optimization procedures have been used (mainly in research settings) to maximize the average admissions test scores and grades of the incoming class, while simultaneously imposing requirements on class composition (e.g., Durán & Wolf-Yadlin, 2011; Kreiter et al., 2003; Pashley & Thornton, 1999; Zwick, 2020; Zwick et al., 2019; Zwick, 2023).

Finally, test-optional and test-blind admissions programs have been growing in popularity (see Zwick, 2019b). The rationale typically provided for moving to a test-optional policy is that it is expected to increase the diversity of the admitted class. Actual results have been mixed.

SUMMARY AND CONCLUSION

Despite the breadth and complexity of the concept of fairness in testing, it may be possible, at least within the measurement community, to agree on a set of general fairness principles:

- Test takers should be treated respectfully and should be fully informed about the purpose of the test and the ways the scores will be used.
- The test, including its content as well as its administration and scoring procedures, should assess the intended construct; no more, no less.
- Tests should “work the same way” for all groups. More formally, measurement invariance, prediction invariance, and linking invariance should hold. For example, equally skilled students should have equal probabilities of answering a test item correctly, regardless of group membership.
- In determining student promotion or graduation, achievement test results should be excluded from consideration unless all students have had the opportunity to learn the tested material.

Of course, these ideals cannot be met perfectly. Consider an example involving one type of measurement invariance—the absence of DIF. We cannot realistically expect that the hypothesis in Equation 3, representing the identity of item response functions across groups, will be perfectly satisfied for every test item and every pair of groups. The unavoidable challenge in implementing these fairness principles is determining how much of a departure from perfect equality is acceptable.

Because the idea of fairness is far too complex to boil down to a single simple definition, fairness claims about a particular test require much more than a set of statistical analysis results or a brief procedural description. As suggested by Zieky (2016), a fairness argument can be crafted as a component of a larger validity argument. In explaining the concept of argument-based validation, Kane (2006) noted that “the proposed interpretations and uses [of test scores are to be] explicitly stated as an argument, or network of inferences and supporting assumptions, leading from observations to the conclusions and decisions” (p. 17). This argument, once publicized, can be evaluated in terms of its coherence and plausibility.

The purpose of a fairness argument is to “build a chain of reasoning using documented evidence to support the fairness of the intended uses and interpretations of the test scores” (Zieky, 2016, p. 96). The argument should first present the specific claims that are being made about the test’s fairness (e.g., “Appropriate accommodations were provided to test takers when needed, in compliance with fairness guidelines”) and should then lay out the pertinent evidence. Support for fairness claims could include information about the quality of the fairness guidelines used in developing and evaluating the test, the training and experience of the test developers and fairness reviewers, the nature of the test-taker groups used in pilot testing, and the results of any relevant statistical analyses, such as assessments of DIF and differential prediction. A fairness argument could address claims about opportunity to learn as well, by including course descriptions, curricula, and classroom observations that document that the targeted material was in fact, taught.

Ideally, a fairness argument should also weigh the effects of test use against the effects of other methods of assessment or selection. In his treatise on validity, Messick (1989) suggested that the evaluation of testing consequences should include a comparison of

the proposed test use to alternative assessment techniques and policies. In the college admissions setting, for example, this comparison would need to include a consideration of both the advantages and the drawbacks of using high school grades and personal characteristics in admissions decisions (see Zwick, 2017b; 2019b). High school grades reflect 4 years of academic accomplishments and provide indirect measures of characteristics like industriousness and perseverance (Willingham et al., 2002), but grading stringency varies over high schools, courses, and teachers, and the assignment of grades can be inappropriately influenced by factors such as a student's SES, gender, or ethnicity. Personal characteristics like conscientiousness and grit seem relevant to college success, and assessments of these characteristics often display smaller group differences than test scores. However, measuring noncognitive factors accurately and without bias is challenging (Duckworth & Yeager, 2015) and studies have raised doubts about their predictive value (Niessen & Meijer, 2020). In general, alternative admissions criteria have lower reliability than standardized tests and have not been studied as thoroughly. In keeping with Messick's recommendation, the research findings about alternative admissions criteria would be considered in weighing the fairness of admissions testing.

Messick noted that "a particularly powerful and general form of counterproposal is to weigh the potential social consequences of the proposed test use against the potential social consequences of *not* testing at all" (1989, p. 86). This type of consideration, in fact, has contributed to the increase in test-optional college admissions policies, which are typically adopted in the hope that they will facilitate campus diversity. Shepard sees the admissions testing situation this way:

If the correlation [of admissions test scores] with success at Berkeley were .7, then doing without the test information might be unfair to high scorers. But given [the much lower actual correlation] and adverse impact, why NOT do without the test? (Lorrie Shepard, personal communication, August 15, 2020)

Arguing from the opposite perspective, Ebel (1966), who is cited by Messick in this context, speculated that the abandonment of educational tests would lead to dire consequences:

the distinctions between competence and incompetence would become more difficult to discern . . . the encouragement and reward of individual efforts to learn would be made more difficult. . . . Educational opportunities would be extended less on the basis of aptitude and merit and more on the basis of ancestry and influence. (p. 28)

Ebel's views echoed the assertions that were made about standardized admissions tests in the early 20th century. The tests, it was claimed, would lead to the admission of the most talented candidates rather than those from the wealthiest or most influential families.

In the early 21st century, institutions that elect to require admissions test scores do so for a number of practical reasons, which I have summarized in Zwick (2019b):

In predicting college performance, using test scores along with high school grades leads to more accurate results than using grades alone. . . . In addition to their utility in predicting college outcomes, standardized admissions tests allow efficient comparisons of students whose high school grades are influenced by the schools they attended and countless extraneous factors. Test scores can provide a way to evaluate international students, home-schooled candidates, and those with a high school equivalency diploma . . . and can sometimes serve to identify talented students with weak academic records (p. 144).

The never-ending debate about college admissions is an illustration of the fact that there will never be consensus on the definition of test fairness. Any mechanism that plays a role in the distribution of desirable goods will always be the subject of contention. In alluding to the multiple meanings of fairness, the *Standards* notes that “it is possible that individuals endorse fairness in testing as a desirable social goal, yet reach quite different conclusions about the fairness of a given testing program” (AERA et al., 2014, p. 49).

While controversies about test fairness will inevitably persist, they should not be permitted to serve as justification for giving short shrift to fairness issues. Given the sharp inequities in our society, it is essential that test makers and test users work toward promoting tests and testing policies that contribute to the fairness of measurement and selection.

ACKNOWLEDGMENTS

I appreciate the thoughtful reviews provided by Linda Cook, Neil Dorans, Mary Pitoniak, Lorrie Shepard, Frank Worrell, and Michael Zieky.

REFERENCES

- Agrawal, N. (2020, October 23). Testing company ACT flagged students' disabilities to colleges. Now it must pay \$16 million. *Los Angeles Times*. <https://www.latimes.com/california/story/2020-10-23/act-settles-lawsuit>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. American Psychological Association.
- Anderson, N. (2020, June 2). College Board suspends plan for at-home SAT, urges colleges not to punish applicants who can't submit scores. *Washington Post*. <https://www.washingtonpost.com/education/2020/06/02/online-sats-coronavirus/>
- Angoff, W. H., & Sharon, A. T. (1974). The evaluation of differences in test performance of two or more groups. *Educational and Psychological Measurement*, 34, 807–816.
- Anrig, G. R. (1987a, January). “Golden Rule.” *Second thoughts*. *APA Monitor*, p. 3.

- Anrig, G. R. (1987b). ETS on "Golden Rule." *Educational Measurement: Issues and Practice*, 6(3), 24–27.
- Association of American Medical Colleges v. Carey, 728 F. Supp. 873 (N.D.N.Y. 1990).
- Atkinson, R. (2001, February 18). *Standardized tests and access to American universities* (The 2001 Robert H. Atwell Distinguished Lecture). American Council on Education Annual Meeting, Washington, DC, United States.
- Bastedo, M. N., Bell, D., Howell, J. S., Hsu, J., Hurwitz, M., Perfetto, G., & Welch, M. (2021). Admitting students in context: Field experiments on information dashboards in college admissions. *The Journal of Higher Education*, 93(3), 327–374. <https://doi.org/10.1080/00221546.2021.1971488>
- Betts, J. R., Rueben, K. S., & Danenberg, A. (2000). *Equal resources, equal outcomes? The distribution of school resources and student achievement in California*. Public Policy Institute of California.
- Blatter, A., & Glynn, J. (2016, November 9–12). *Recognizing distance traveled in selective college admissions* [Paper presentation]. Association for the Study of Higher Education Annual Meeting, Columbus, OH, United States.
- Breimhorst v. Educational Testing Service, No. C-99-3387 WHO (N.D. Cal. 2000).
- Brigham, C. C. (1923). *A study of American intelligence*. Princeton University Press.
- Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review*, 37, 158–165.
- Camara, W. J. (2009). College admission testing: Myths and realities in an age of admissions hype. In R. P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147–180). American Psychological Association.
- Camara, W. J., & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification* (College Board Report 99-5). College Entrance Examination Board.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). ACE/Praeger.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage.
- Carnevale, A. P., & Haghigat, E. (1998). Selecting the strivers: A report on the preliminary results of the ETS "educational strivers" study. In D. Bakst (Ed.), *Hopwood, Bakke, and beyond: Diversity on our nation's campuses* (pp. 122–128). AACRAO.
- Chin, M. (2020, December 9). Privacy group files complaint against five online test-proctoring services. *The Verge*. <https://www.theverge.com/2020/12/9/22166023/epic-proctorio-examity-privacy-online-testing-school-lawsuit-proctoring>
- Chiorri, C., Marsh, H. W., Ubbiali, A., & Donati, D. (2016). Testing the factor structure and measurement invariance across gender of the Big Five Inventory through exploratory structural equation modeling. *Journal of Personality Assessment*, 98(1), 88–99.
- Civil Rights Act. (1991). Title VII 703(a)(1), 42 U.S.C. 2000-2(a) (1), as amended by CRA of 1991, section 106.

- Class Action Lawsuit Filed Against the College Entrance Examination Board Over Advanced Placement Testing*. (2020, May 27). Justia. <https://news.justia.com/class-action-lawsuit-filed-against-the-college-entrance-examination-board-over-advanced-placement-testing/>
- Cleary, T. A. (1968). Test bias: Prediction of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5(2), 115–124.
- Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, 10, 237–255.
- Cole, N. S. (1981). Bias in testing. *American Psychologist*, 36, 1067–1077.
- Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25–29). Lawrence Erlbaum Associates.
- Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). American Council on Education/Macmillan.
- Cronbach, L. J. (1975). Five decades of controversy over mental testing. *American Psychologist*, 30, 1–14.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.
- Culpepper, S. A., Aguinis, H., Kern, J. L., & Millsap, R. (2019). High-stakes testing case study: A latent variable approach for assessing measurement and prediction invariance. *Psychometrika*, 84(1), 285–309.
- Darlington, R. B. (1971). Another look at “cultural fairness.” *Journal of Educational Measurement*, 8, 71–82.
- Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla 1979). <https://www.courtlistener.com/flmd/b52Q/debra-p-v-turlington/>
- Debra P. v. Turlington, 730 F.2d 1405 (1984). <http://law.justia.com/cases/federal/appellate-courts/F2/730/1405/345196/>
- Doe v. National Board of Medical Examiners, No. 99-4532 (E.D. Pa. 1999).
- Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41, 43–68.
- Dorans, N. J., & Cook, L. (2016). The implications of societal changes for fairness assessment. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 295–297). Routledge.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281–306.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.
- Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, 49, 1–18.
- Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, 44, 237–251.

- Durán, G., & Wolf-Yadlin, R. (2011). A mathematical programming approach to applicant selection for a degree program based on affirmative action. *Interfaces*, 41(3), 278–288.
- Ebel, R. L. (1966). The social consequences of educational testing. In A. Anastasi (Ed.), *Testing problems in perspective* (pp. 18–28). American Council on Education.
- ETS. (2022). *ETS guidelines for developing fair tests and communications*. <https://www.ets.org/content/dam/ets-org/pdfs/about/fair-tests-and-communications.pdf>
- Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, 75, 261–269.
- Equal Employment Opportunity Program. (2021). *National archives*. <https://www.archives.gov/eo/terminology.html>
- Faggen, J. (1987). Golden Rule revisited: Introduction. *Educational Measurement: Issues and Practice*, 6(2), 5–8.
- Gaertner, M. N., & Hart, M. (2013). Considering class: College access and diversity. *Harvard Law and Policy Review*, 7, 367–403.
- Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, 8, 1–26.
- Gewertz, C. (2018, October 4). ACT stops “flagging” scores of students with disabilities. *Education Week*. <https://www.edweek.org/teaching-learning/act-stops-flagging-scores-of-students-with-disabilities/2018/10>
- Goggin, W. J. (1999, May). A “merit-aware” model for college admissions and affirmative action. *Postsecondary Education Opportunity Newsletter*, pp. 6–12.
- Golden Rule Insurance Company v. Washburn, No. 419-76 (Ill. Cir. Ct. Nov. 20, 1984).
- Gould, S. J. (1996). *The mismeasure of man* (Revised & expanded ed.). Norton.
- The Graide Network. (2018, August 27). *Teacher bias: The elephant in the classroom*. <https://www.thegraidenetwork.com/blog-all/2018/8/1/teacher-bias-the-elephant-in-the-classroom>
- Gross, A. L., & Su, W. (1975). Defining a “fair” or “unbiased” selection model: A question of utilities. *Journal of Applied Psychology*, 60, 345–351.
- Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher*, 32(5), 19–25.
- Hackett, R. K., Holland, P. W., Pearlman, M., & Thayer, D. T. (1987). *Test construction manipulating score differences between Black and White examinees: Properties of the resulting tests* (ETS Research Report No. 87-30). ETS.
- Herold, B. (2020, April 10). The disparities in remote learning under coronavirus (in charts). *Education Week*. <https://www.edweek.org/ew/articles/2020/04/10/the-disparities-in-remote-learning-under-coronavirus.html>
- Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics*. Macmillan.
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Erlbaum.

- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Hoover, E. (2019, October 29). U. of California faces bias lawsuit over ACT/SAT requirement. *Chronicle of Higher Education*. <https://www.chronicle.com/article/U-of-California-Faces-Bias/247434>
- Hoover, E. (2020, September 1). Court bars U. of California from using ACT and SAT for fall-2021 admissions. *Chronicle of Higher Education*. <https://www.chronicle.com/article/court-bars-u-of-california-from-using-act-and-sat-for-fall-2021-admissions>
- Hu, J. C. (2020, June 24). Graduate programs drop GRE after online version raises concerns about fairness. *Science*. <https://www.sciencemag.org/careers/2020/06/graduate-programs-drop-gre-after-online-version-raises-concerns-about-fairness>
- Hughes, T. (1998, February 18). *Admissions criteria: Standardized testing* (California State Senate Bill 1807). Legislative Counsel's Digest. <http://www.sen.ca.gov>
- Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology*, 3, 131–150.
- Jaschik, S. (2020a, August 27). NACAC urges public colleges to go test optional. *Inside Higher Ed*. <https://www.insidehighered.com/quicktakes/2020/08/27/nacac-urges-public-colleges-go-test-optional>
- Jaschik, S. (2020b, November 2). Appeals court upholds ruling against tests for admissions. *Inside Higher Ed*. <https://www.insidehighered.com/admissions/article/2020/11/02/appeals-court-upholds-ruling-barring-use-sat-and-act-u-california>
- Jensen, A. R. (1980). *Bias in mental testing*. Free Press.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. <https://www.apa.org/science/programs/testing/fair-testing.pdf>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). ACE/Praeger.
- Kane, M., & Bridgeman, B. (2021). The evolution of the concept of validity. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 181–205). Routledge.
- Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, 41, 3–14.
- Kreiter, C. D., Stansfield, B., James, P. A., & Solow, C. (2003). A model for diversity in admissions: A review of issues and methods and an experimental approach. *Teaching and Learning in Medicine: An International Journal*, 15(2), 116–122.
- Ladson-Billings, G. (1995). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32, 465–491.
- Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice*, 39(3), 24–27.
- Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Economic Policy Institute.
- Linn, R. L. (1973). Fair use in selection. *Review of Educational Research*, 43, 139–161.

- Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, 21, 33–47.
- Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
- Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, 7, 29–38.
- Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement: Issues and Practice*, 6(2), 13–17.
- Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, 8, 1–4.
- Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice*, 32(1), 15–32.
- Liu, J., & Dorans, N. J. (2016). Fairness in score interpretation. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 77–96). Routledge.
- Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Swets & Zeitlinger.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.
- Manning, W. H. (1977). *The pursuit of fairness in admissions to higher education*. In *Selective admissions in higher education* (Carnegie Council on Policy Studies in Higher Education report, pp. 19–64). Jossey-Bass.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Map: Coronavirus and School Closures in 2019–2020. (2020, March 6). *Education Week*. <https://www.edweek.org/ew/section/multimedia/map-coronavirus-and-school-closures.html>
- Marco, G. L. (1988). Does the use of test assembly procedures proposed in legislation make any difference in test properties and in the test performance of Black and White test takers? *Applied Measurement in Education*, 1(2), 109–133.
- McAllister, P. H. (1991). Overview of state legislation to regulate standardized testing. *Educational Measurement: Issues and Practice*, 10(4), 19–22.
- Medina, J., Benner, K., & Taylor, K. (2019, March 12). Actresses, business leaders and other wealthy parents charged in U.S. college entry fraud. *New York Times*. <https://www.nytimes.com/2019/03/12/us/college-admissions-cheating-scandal.html>
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/ Macmillan.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

- Miyazaki, I. (1976). *China's examination hell: The civil service examinations of imperial China*. Weatherhill.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
- National Center for Education Statistics. (2022). Characteristics of children's families. In *Condition of education* (pp. 1–10). U.S. Department of Education, Institute of Education Sciences. <https://nces.ed.gov/programs/coe/indicator/cce>.
- National Center on Educational Outcomes. (2016). *Universal design of assessments: Overview*. https://nceo.info/Assessments/universal_design
- Niessen, A. S. M., & Meijer, R. R. (2020). Character-based admissions criteria in the United States and in Europe. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: An international perspective* (pp. 76–95). Cambridge University Press.
- Novick, M. R., & Ellis, D. D. (1977). Equal opportunity in educational and employment selection. *American Psychologist*, 32, 306–320.
- Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. U.S. Government Printing Office.
- Organisation for Economic Co-operation and Development (OECD, 2010). *PISA 2009 results: Overcoming social background – equity in learning opportunities and outcomes* (Volume II). OECD Publishing. <http://dx.doi.org/10.1787/9789264091504-en>
- Owings, J., McMillen, M., & Burkett, J. (1995). *Making the cut: Who meets highly selective college entrance criteria?* (NCES 95-732). National Center for Education Statistics. <http://nces.ed.gov/pubs95/95732.pdf>
- Pashley, P. J., & Thornton, A. E. (1999). *Crafting an incoming law school class: Preliminary results*. (LSAC Research Report 99–01). Law School Admission Council.
- Patil, A., & Bromwich, J. E. (2020, September 29). How it feels when software watches you take tests. *New York Times*. <https://www.nytimes.com/2020/09/29/style/testing-schools-proctorio.html?referringSource=articleShare>
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C. R. Rao (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 125–167). Elsevier.
- Petersen, N. S. (1976). An expected utility model for “optimal” selection. *Journal of Educational Statistics*, 1, 333–358.
- Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, 13, 3–29.
- Phillips, S. E. (2016). Legal aspects of test fairness. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 239–266). Routledge.
- Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME *Standards for Educational and Psychological Testing*? *Educational Measurement: Issues and Practice*, 33(4), 4–12.
- Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13.

- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367–388). Lawrence Erlbaum Associates.
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.
- Rauschenberg, S. (2014). How consistent are course grades? An examination of differential grading. *Education Policy Analysis Archives*, 22(92). <http://dx.doi.org/10.14507/epaa.v22n92.2014>
- Ravitch, D. (2003, May 11). Excerpt from *The Language Police*. *New York Times*. <https://www.nytimes.com/2003/05/11/books/chapters/the-language-police.html>
- Rebell, M. A. (1989). Testing, public policy, and the courts. In B. R. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 135–162). Kluwer Academic.
- Rothstein, J. M. (2004). College performance and the SAT. *Journal of Econometrics*, 121, 297–317.
- Sackett, P. R., & Kuncel, N. R. (2018). Eight myths about standardized admissions testing. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Testing, academic achievement, and the future of college admissions* (pp. 13–39). Johns Hopkins University Press.
- Saretzky, G. D. (1982). *Carl Campbell Brigham, the native intelligence hypothesis, and the Scholastic Aptitude Test* (ETS Research Memorandum No. RR-82-4). ETS.
- Schmitt, A. P., Curley, W. E., Bleistein, C. A., & Dorans, N. J. (1988, April 5–9). *Experimental evaluation of language and interest factors related to differential item functioning for Hispanic examinees on the SAT-Verbal* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, United States.
- Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Lawrence Erlbaum Associates.
- Settlement Agreement and Release of All Claims*. (2021). Public Counsel. https://publiccounsel.org/wp-content/uploads/2021/12/Smith-v-UC-Regents_SETTLEMENT-AGREEMENT.pdf
- Shealy, R. T., & Stout, W. F. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 197–239.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–8, 13, 24.
- Shepard, L. A. (2012). Equity in assessment. In J. A. Banks (Ed.), *Encyclopedia of diversity in education* (Vol. 2, pp. 807–812). SAGE.
- Sireci, S., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 111–134). Routledge.

- Smith v. Regents of University of California. (2019, December 10). *Alameda Superior Court case number RG19046222*. Public Counsel. <https://edsources.org/wp-content/uploads/2019/12/2019.12.10-Kawika-Smith-et-al-v.-UC-Regents-et-al.pdf>
- Stricker, L. J., Rock, D. A., Pollack, J. M., & Wenglinsky, H. H. (2002). *Measuring educational disadvantage of SAT candidates* (College Board Research Report No. 2002-1). College Entrance Examination Board.
- Students Sue to Enjoin UC Regents' Use of False "Test-Optional" Admissions Policy for SAT and ACT*. (2020). [Press release, Public Counsel]. http://www.publiccounsel.org/press_releases?id=0135
- Studley, R. E. (2004). Inequality, student achievement, and college admissions: A remedy for underrepresentation. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 321–343). RoutledgeFalmer.
- Swaak, T. (2022, August 24). Students say room scans during online tests are invasive. Now a judge agrees. *Chronicle of Higher Education*. https://www.chronicle.com/article/students-say-room-scans-during-online-tests-are-invasive-now-a-judge-agrees?cid=gen_sign_in
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet–Simon Intelligence Scale*. Houghton Mifflin.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates.
- Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, 8, 63–70.
- University of California Academic Senate. (2020, January). *Report of the UC Academic Council Standardized Testing Task Force*. https://senate.universityofcalifornia.edu/_files/committees/sttf/sttf-report.pdf
- U.S. Department of Education. (2017, December 7). *Every Student Succeeds Act assessments under Title I, part A & Title I, part B: Summary of final regulations*. <https://www2.ed.gov/policy/elsec/leg/essa/essaassessmentfactsheet1207.pdf>
- Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30, 1045–1058.
- Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles* (ETS Research Report No. RR-23-11). ETS. <https://doi.org/10.1002/ets2.12374>
- Watanabe, T. (2021, November 18). UC slams the door on standardized admissions tests, nixing any SAT alternative. *Los Angeles Times*. <https://www.latimes.com/california/story/2021-11-18/uc-slams-door-on-sat-and-all-standardized-admissions-tests>

- Webber, C. (1989). The mandarin mentality: University admissions testing in Europe and Asia. In B. R. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 33–57). Kluwer Academic.
- Whipple, G. M. (1921). The National Intelligence Tests. *Journal of Educational Research*, 4, 16–31.
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39, 1–37.
- Wilner, J. (2020, May 21). University of California eliminates SAT, ACT exams from admissions process throughout the system. *San Jose Mercury News*. <https://www.mercurynews.com/2020/05/21/university-of-california-eliminates-sat-act-exams-from-admissions-process-throughout-the-system/>
- Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–301). RoutledgeFalmer.
- Young, J. W., & Johnson, P. M. (2004). The impact of an SES-based model on a college's undergraduate admissions outcomes. *Research in Higher Education*, 45, 777–797.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Lawrence Erlbaum Associates.
- Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.
- Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, 15, 185–197.
- Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. RoutledgeFalmer.
- Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report. No. RR-12-08). ETS.
- Zwick, R. (2017a). *Who gets in? Strategies for fair and effective college admissions*. Harvard University Press.
- Zwick, R. (2017b, July). The risks of focusing on character in admissions. *Chronicle of Higher Education*. <http://www.chronicle.com/article/The-Risks-of-Focusing-on/240787?cid=wsinglestory>
- Zwick, R. (2019a). Fairness in measurement and selection: Statistical, philosophical, and public perspectives. *Educational Measurement: Issues and Practice*, 38(4), 34–41.
- Zwick, R. (2019b). Assessment in American higher education: The role of admissions tests. *Annals of the American Academy of Political and Social Science*, 683, 130–148.
- Zwick, R. (2020). Using mathematical models to improve access to postsecondary education. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: An international perspective* (pp. 333–346). Cambridge University Press.
- Zwick, R. (2021). A century of testing controversies. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 136–154). Routledge.

- Zwick, R. (2023). Using quantitative techniques to promote diversity in admissions. In K. F. Geisinger (Ed.), *College admissions and admissions testing in a time of transformational change*, pp. 151–164. Routledge.
- Zwick, R., Blatter, A., Ye, L., & Isham, S. (2021). Using constrained optimization with an index of admission obstacles to increase the diversity of college classes, *Educational Assessment*, 26, 20–34.
- Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing* (pp. 271–284). CRC Press.
- Zwick, R., & Dorans, N. J. (2016). Philosophical perspectives on fairness in educational assessment. In N. J. Dorans & L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 267–281). Routledge.
- Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, 26, 55–66.
- Zwick, R., & Green, J. G. (2007). New perspectives on the correlation of SAT scores, high school grades, and socioeconomic factors. *Journal of Educational Measurement*, 44, 23–45.
- Zwick, R., & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement*, 48, 101–121.
- Zwick, R., & Sklar, J. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42, 439–464.
- Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing DIF in polytomous items. *Applied Measurement in Education*, 10, 321–344.
- Zwick, R., Thayer, D. T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement*, 18, 121–140.
- Zwick, R., Ye, L., & Isham, S. (2019). Using constrained optimization to increase the representation of students from low-income neighborhoods. *Applied Measurement in Education*, 32, 281–297.

NOTES

1. There is some disagreement on this point. Some historians include tests given much earlier, during the Han dynasty (206 B.C.E.–220 CE; Webber, 1989).
2. The statistical meaning of bias is not related in any way to fairness. “Any statistic whose mathematical expectation is equal to a parameter $[\mu]$ is called an *unbiased* estimator of the parameter $[\mu]$ ” (Hogg & Craig, 1978, p. 204). Suppose that in a particular population—say eighth graders in the United States—the average score on a test is μ . We can estimate μ by taking a random sample of eighth graders and calculating the sample mean of their test scores, $\bar{X}$. We know from statistical

theory that $E(\bar{X}) = \mu$; that is, the expectation of $\bar{X}$, which is its theoretical average over random samples, is equal to μ . Therefore, we can say that $\bar{X}$ is an unbiased estimate of μ . This tells us nothing whatever about whether the test in question is a fair test.

3. From a practical standpoint, it is necessary to limit the examination of consequences to some degree. For example, it would be neither possible nor desirable to attempt to determine all possible effects of a given selection test throughout the lifetimes of the test takers. The term *proximal consequences* may be useful, although it does not resolve the ambiguity of meaning.
4. The use of the term *construct* in the *Standards* differs from some historical uses. Rather than reserving the term for characteristics that are not directly observable, the *Standards* uses *construct* to refer to “the concept or characteristic that a test is designed to measure” (AERA et al., 2014, p. 11). I adhere to this usage throughout this chapter.
5. For example, in 2022, the National Council on Measurement in Education held a webinar titled *Cultural Relevance Versus Construct Relevance*. The description includes the following: “Traditional assessment practices have suggested that to make assessments accessible to all students, regardless of racial or cultural background, we must remove any construct-irrelevant variable, including culture-specific contextual items, that may unintentionally privilege or hinder performance. However, a growing body of literature calls for the use of culture-specific contextual items, citing that culturally relevant material will increase engagement, interest, and inclusion of marginalized students.” See <https://www.youtube.com/watch?v=bB1FtEJLqAE>. For further discussion, see Walker et al. (2023).
6. A lexicon of Equal Employment Opportunity Program Terminology appearing on a U.S. government website lists the following groups as having “minority” status because of their exclusion from the economy: American Indian or Alaskan Native, Asian or Pacific Islander, Black, and Hispanic. It is also noted that “women are not classified as a minority. However, they have experienced the same kind of systematic exclusion from the economy as the various minorities. Thus, they are considered as having ‘minority status’ as far as the law is concerned” (Equal Employment Opportunity Program, 2021). It should be noted that this lexicon is rendered obsolete by the January 2025 rescission of the Equal Employment Opportunity Act of 1965.
7. Because the DIF analysis for a single test typically involves a large number of hypothesis tests, it is sometimes suggested that multiple comparison procedures, such as the Bonferroni method, be used to control the overall Type I error rate. The opposing point of view, which has largely prevailed, is that Type II errors are of greater concern in the DIF context, implying that the usual concerns about the compounding of Type I error are inappropriate in this case. Surprisingly little discussion has been devoted to this issue.

8. My conceptualization of DIF departs from that of Liu and Dorans (2016, pp. 84–85) in that they view DIF methods based on observed scores as ways of assessing prediction invariance, whereas DIF methods that invoke latent variable models are assumed to assess measurement invariance. I consider all DIF methods to be ways of assessing measurement invariance, regardless of whether the analyses explicitly involve latent variables. The distinction between DIF methods based on IRT models and those based on observed scores is somewhat artificial because the IRT methods must rely on *estimates* of θ (or the distribution of θ) that are themselves based on the observed item scores.
9. Variations of these methods that use scores derived through IRT (e.g., see Zwick et al., 1994) or other alternative scoring techniques have been developed.
10. Informally speaking, sufficiency of S in this case means that the distribution of the item scores, conditional on S , does not depend on θ . That is, θ provides no additional information about the item scores beyond what is contained in S .
11. If $p = 1$ and X and Y have been standardized to have mean zero and variance 1, then a test of differential prediction via regression (see Equation 11) is equivalent to a test of differential validity. This equivalence does not hold for $p > 1$.

Part III

Developing, Administering, and Scoring Assessments

Designing and Developing Educational Assessments

Kristen Huff

Curriculum Associates

Paul Nichols

Retired

M. Christina Schneider

Cambium Assessment

The intended audiences for this chapter are students and professionals—both early career and seasoned—engaged in the design and development of educational assessments. The focus of this chapter is the design and development of large-scale interim and summative assessments of educational achievement for which the primary intended purposes are reliable classification of student performance into one of three or more categories, or performance levels, with the intended inferences of what students know and can do with regard to specified learning standards supported by a compelling validity argument. For the purposes of this chapter, we define interim assessments as those that occur multiple times throughout a school year and are used primarily to inform instruction and track student progress. The term summative is used to refer to assessments that occur less frequently, such as once per year, and are used primarily for educational accountability. We believe that the design of educational interim and summative assessments as defined here and with these respective purposes is the same.¹

Since the rise and proliferation of federally mandated end-of-year testing began in 1994 (Improving America's Schools Act, 1994), the educational measurement field has learned a great deal about what is needed to design and develop assessments that are intended to support interpretations about what students know and can do (Ferrara & DeMauro, 2006; Lane et al., 2016; Mislevy, 2006; National Research Council [NRC], 2001; Pellegrino et al., 1999, 2016; Rupp & Leighton, 2017). However, many of the recommendations by these scholars and practitioners are still, frustratingly, emergent rather than commonplace (Ferrara & DeMauro, 2006; Huff et al., 2017). Supporting valid inferences about what students have learned requires more sophisticated approaches to test design and development than in the past, when most educational tests were designed primarily, if not solely, for various norm-referenced decision-making or rank ordering (e.g., college admissions, to select the top 2% for gifted and talented programs, or the bottom 20% for special services). We posit that contemporary educational testing is faced with additional complex challenges that routine design practices simply cannot meet; we group these challenges into four categories: (a) the complexity of target constructs, (b) using assessment results for multiple purposes, (c) assessment quality, and (d) accessibility for increasingly diverse student populations.

First, the constructs of interest are being articulated in more sophisticated ways than in the past, as our scholarship continues to evolve in the learning sciences and as learning standards begin to reflect the sociocognitive nature of learning: how students develop the kinds of deep understanding required not only for disciplinary practice but also that set the stage for additional learning (Mislevy, 2006; NRC, 2000, 2001, 2002, 2005; Penuel & Shepard, 2017). The following are examples of standards and frameworks that reflect increasing complexity than their respective prior versions: The Common Core State Standards (CCSS; National Governors Association Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), the National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy Framework (National Assessment Governing Board [NAGB], 2018); the 2026 NAEP

Mathematics and Reading Frameworks (NAGB, 2021a, 2021b) and the Advanced Placement curriculum frameworks for science and history (College Board, 2019).

For example, as of this writing, most states have based their K–12 reading learning standards on the CCSS, which set expectations for comprehension from multiple sources, such as reading Martin Luther King’s “Letter from a Birmingham Jail” and listening to one of his speeches and then responding to a writing prompt by citing evidence from both sources. In mathematics, there is now an expectation that students go beyond procedural fluency to conceptually understand multiple ways to problem-solve as an aspect of demonstrating proficiency in mathematical practices. Another example is the multidimensional nature of the NGSS, which integrate concepts that span the science domains, discipline-specific knowledge, and scientific practices. Therefore, the tasks we devise to measure these complex learning standards with unprecedented cognitive load, the coding schemes we use to score student responses, the measurement models we use to estimate student proficiency, and the concepts and indices used to evaluate technical quality of the assessment all need to reflect the intended complexity of the construct of interest. As an example of the new demands on evaluating technical quality, the conventional notions of comparability and reliability will likely need to evolve to keep up with contemporary needs.

Second, in addition to the challenges of measuring student proficiency with regard to sophisticated constructs, test users are demanding that educational assessments, both interim and summative, serve a variety of purposes as all manner of stakeholders—policy makers, parents, teachers, and students themselves—call for less testing, which results in the need for a greater variety of uses of the assessments that remain in place (Hart et al., 2015; Huff & Goodman, 2007). In practice, the results of interim and summative assessments are used to inform determinations at various levels of the educational system, including but not limited to instructional next steps at the individual or group level, whether the student needs additional testing for English language proficiency or dyslexia, grade promotion, algebraic readiness, educator effectiveness, school ratings, and district- and state-level policy decisions about curriculum and professional development resources. According to the *2014 Standards for Educational and Psychological Testing* (see Chapter 1; American Educational Research Association [AERA] et al., 2014), it is incumbent on the user to collect and evaluate evidence to support the assertion that the inferences from the assessment are valid for any use that is additional to or a departure from the purposes and use for which the assessment was originally designed and, presumably, validated (see also Lane & Marion, this volume). To meet the demands of less testing but increased usage of the results of any single test, test providers are looking for ways to design assessments from the beginning to serve more than one purpose.

With increased public discourse about the role of educational testing in schools, a third challenge has emerged that is not likely to subside anytime soon: Test quality is under scrutiny. As a condition for some federal education funding, the United States Department of Education requires state summative assessments to be reviewed for

technical quality by a panel of peers. Similarly, some states are requiring reviews for interim assessments before they are approved for use in classrooms (e.g., Louisiana Department of Education, 2020; South Carolina Department of Education, n.d.). As the level of scrutiny increases, there is greater burden on the test makers to have documented, transparent, and understandable validation arguments that support the variety of intended inferences about what students know and can do.

The fourth and last category of contemporary assessment design challenges is the increased need to ensure that our assessments provide results that support valid inferences about all students regardless of their sociocultural background, including home language, or disability.² We refer to this aim as ensuring that our assessments are accessible to all students. Teaching and learning are fundamentally sociocultural endeavors (Penuel & Shepard, 2017) that need to acknowledge and build on students' background knowledge and cultural experience. Assessment needs to reflect this principle to remain authentic and relevant to students' classroom experience. Additionally, there is a growing research base exploring the hypothesis that without this contextual relevance, we will not engage students during the assessment, which will demotivate them and thus undermine the accuracy of the resulting scores as a measurement of best performance, in part because of lack of consideration of construct-irrelevant factors that are present when sociocultural factors are not considered (Brown et al., 1989; Wise, 2020).

Students with disabilities must also have equal opportunity to demonstrate what they know and can do through accessible assessments. Implementation of the universal design for learning principles that guide assessment design (CAST, 2018; Johnstone et al., 2006) have been available to inform practice for some time for both paper-and-pencil and computer-based assessments. In addition, as more assessment systems are delivered via the Internet, many providers hold themselves accountable to the Web Content Accessibility Guidelines (WCAG; World Wide Web Consortium (W3C), 2018), which address features such as use of color and images, speech-to-text availability, captioning, and more. Simultaneously ensuring that students from various sociocultural backgrounds are engaged during the assessment, that the assessment is free of barriers for students with disabilities and multilanguage learners, and that the target of measurement has not been compromised requires a more contemporary approach to assessment design than is currently commonplace. When the targets of measurement are defined such that all ambiguity is eliminated, then discussions about what constitutes construct-relevant versus construct-irrelevant variance can occur with a level of precision that is typically absent, allowing for designing assessments that eliminate barriers to access that might otherwise be present. Creating assessments that are free of bias, are accessible to all students, and are culturally and linguistically responsive requires undisputed clarity on the target of measurement by the full interdisciplinary team.

It is in the context of these contemporary challenges that we write this chapter. To meet contemporary demands, educational achievement tests must be designed with more deliberate, interdisciplinary decision-making about each design feature and must make all assumptions and rationales transparent for interrogation. The intended inferences

about what students know and can do must have a compelling—and understandable—validation argument that resonates with policy makers, educators, and parents. Transparent design and a compelling, understandable validation argument are critical to support the varied intended purposes and uses of educational interim and summative assessments, especially in a climate when educational testing is under unprecedented scrutiny and criticism. In this chapter, we illustrate how principled assessment design (PAD) is not a rigid set of processes that require a new jargon-filled lexicon, but rather a discipline requiring a particular mindset and the use of tools that can help us build assessments that meet contemporary educational needs.

Two assertions are made throughout the chapter. First, assessments designed with conventional approaches do not adequately serve the purposes for which assessments are used in 21st-century educational systems; PAD offers a solution. Given the primary use case for large-scale assessments of educational achievement—to support valid inferences of what students know and can do in relation to a given set of learning outcomes for a given purpose and use—we, as a field, can no longer support approaches to educational assessment design and development that implicitly assume the measurement of a largely fixed latent trait for the purpose of rank ordering and selecting students and that define adequate validation evidence as a disconnected series of post hoc analyses to be collected and published in the technical manual. This is especially true as the constructs to be assessed become more complex, assessment results are used for a variety of purposes, assessment quality is under more public scrutiny, and our understanding of accessibility becomes simultaneously broader and more nuanced.

Conventional approaches to educational assessment design are fragmented and must be integrated and supplemented to best serve the intended role of large-scale educational assessments, whether designed to inform instruction, to make student-level decisions, or for broader accountability purposes. Fragmentation in design and development undermines the coherence required to support the inferences the assessment results are intended to support within the larger context of the educational endeavor (Herman, 2010; Huff et al., 2017). Thus, our second assertion is that PAD integrates what is fragmented in conventional approaches into a coherent design approach and validation argument. For example, historically there has been a stark divide in teams and processes between the item writing and the analysis of the resulting data, as if the creation of the test were an assembly line or a relay with handoffs. One clear consequence of this disjunction is the existence of test content specifications that are separate from the test psychometric specifications. In conventional approaches to assessment design, these two design elements are developed separately by respective teams with the exception of a few touchpoints. This divide in mindset and practice can be bridged through the discipline and practice of PAD.

PAD is an umbrella of test design approaches including construct-centered measurement (Messick, 1994; Wilson, 2005), cognitive design systems (Embretson, 1998), evidence-centered design (ECD; Huff et al., 2010; Mislevy & Haertel, 2006; Mislevy et al., 2003; Pearlman, 2008a, 2008b), principled design for efficacy (Nichols et al., 2016),

and assessment engineering (Luecht, 2013). PAD is distinguished from conventional approaches to assessment design and development by three overarching characteristics (see Figure 7.1).

First, the construct is defined by how students learn and build knowledge in the domain of interest, with a careful consideration that students from different cultural backgrounds bring different funds of knowledge that need to be valued and serve as the foundation for learning (Gay, 2000; Randall, 2021). A compelling example of construct definition is how the Advanced Placement (AP) program redesigned its science and history exams over the course of several years beginning in 2005 in response to criticism from the NRC (2002), where both the AP and International Baccalaureate examinations were criticized for overemphasizing declarative and procedural knowledge and for a lack of pedagogical and learning science research in science and historical thinking to inform assessment design. Given that assessments have consequences for what is taught and valued in the classroom, the NRC challenged the test providers to do better at ensuring that what is measured on the exam is also valued in the discipline. Other examples of this move away from overvaluing declarative and procedural knowledge as targets of measurement are the focus on scientific practices in the NGSS and the focus on mathematical practices in the CCSS. Both sets of learning standards were informed by research from the learning sciences and practitioners in the respective disciplines (CCSSO, 2010; NGSS Lead States, 2013).

A second distinguishing characteristic of PAD is the explicit articulation of all assumptions, design decisions, and rationales for those decisions, with particular attention to our need to serve a diverse student body—students from various sociocultural backgrounds, multilanguage learners, and student with disabilities. The various articulations result in a set of design tools that are used throughout the design and development process and are also subject to continuous improvement throughout the life of the assessment program as data are collected from students and used as feedback to inform refinements to our assumptions, design decisions, and tools. Use of these

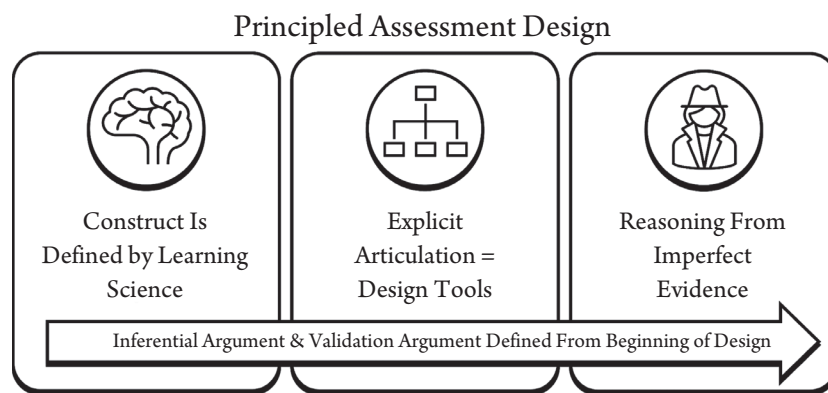


FIGURE 7.1
Distinguishing Characteristics of Principled Assessment Design

tools consistently by all players in the design and development endeavor not only helps to ensure coherence from construct definition through task development and score inference (and all the steps in between), but also serves as infrastructure for the inferential, evidential, and validation arguments. Briefly, the inferential argument is the chain of reasoning that is woven throughout the assessment design (e.g., How do these items measure the intended construct and support the claims about what students know and can do?), the evidential argument is the evidence for the inferential argument, and the validation argument is how well the evidence supports the intended inferences.

Finally, the last distinguishing characteristic of PAD is an adherence to the mindset that assessment is the process of reasoning from imperfect evidence, and that process begins with design. Reasoning from first principles—in short, leaving no assumption left implicit—is an explicit requirement throughout the design, development, and evaluation process.

The benefits of this shared discipline of PAD are many. Employing PAD will force us to make intentional and explicit the necessary relationships among performance-level descriptors (PLDs), task design, and scale properties that result in a strong body of evidence to support inferences about what students know and can do. In other words, we need the psychometric properties of the scale (e.g., sufficient measurement information across the scale, especially at various cut scores) to support reliable classification of students into performance categories and for those classifications to support valid inferences about what students know and can do. Rather than leaving this to happenstance, we need to purposefully design tasks that require students to demonstrate the knowledge and skills associated with various performance categories. PLDs are our guidance for task design at the beginning of assessment design and our guidance for score interpretation once we have scores for each student. That is, the PLDs articulate the inferences we eventually want to make about what students know and can do, and using them as a primary element of task design is a strong foundation for the inferential and validation arguments for the assessment.

In PAD, there is an almost obsessive focus on clarity, articulation, and documentation of all aspects of assessment design, especially the targets of measurement, the evidence required to support inferences about those targets of measurement, and the tasks that are best suited to collect that evidence given the constraints of the assessment. Later in the chapter, we will address how the evidentiary arguments supporting how specific tasks yield valid evidence for inferences regarding specific targets of measurement are a subcomponent of a larger evidentiary argument for the assessment as a whole. Suffice to say for now that these intentionally designed layers of evidence, made explicit and transparent, result in increased clarity in the field about the intended targets of measurement of the assessment and the constructs to which the inferences about students can be generalized, which results in greater coherence between what inferences the assessment results support with regard to what students know and can do and the role the assessment plays in a larger educational context. This clarity and coherence help the assessment results better serve the diverse needs of policy makers, educators, students, and parents—and it is our hope that this, in turn, will make the results more meaningful

for them. As a field, we must make our assessments more meaningful and useful for educators, parents, and students—not only because it is the right thing to do to support teaching and learning, but also in light of the many demands facing educational testing in the early 21st century. PAD gives us a way to do that.

SIMILARITIES TO CONVENTIONAL ASSESSMENT DESIGN AND DEVELOPMENT

In 1989, Jason Millman and Jennifer Greene wrote the test development chapter for *Educational Measurement*, third edition, which opened,

This chapter is about making tests. It is directed to the professional test constructor, not to the classroom teacher. Our goal is to emphasize options for specifying and developing tests, not to produce a procedural manual. (p. 335)

We have the same goal. There are many sources that outline in much detail the precise steps for developing an assessment. Specifically, the *Handbook of Test Development*, second edition (Lane et al., 2016), includes 32 chapters devoted to every aspect of test development, administration, scoring, and evaluating validity evidence for the intended purpose and use of the assessment. The editors did an excellent job of balancing conventional approaches and PAD in their selection of authors and chapter topics. Another excellent resource is the 2013 publication by CCSSO and the Association of Test Publishers entitled *Operational Best Practices for Statewide Large-Scale Assessment Programs*.

Operationalizing PAD does not exempt the test developer from the nuts and bolts of test development; appropriate subject matter experts (e.g., classroom teachers, learning scientists) are engaged in the process, items are developed and reviewed against various criteria, items are field tested, and psychometric analyses are conducted (including but not limited to analysis of item difficulty, discrimination, differential item functioning, scaling, equating, standard setting, and more). Rather than devote this chapter to an explication of these well-documented processes and procedures, we have attempted—like many before us since the early 1980s—to question the status quo and make a compelling argument for shifting our practice to better support inferences about what students know and can do. As Huff et al. (2010) said of ECD—which applies equally to PAD—“ECD can be a first step toward challenging the assumptions and ‘breaking out of the current paradigm’ of large-scale assessment” (p. 316).

THE EVOLUTION OF PAD

There is a saying, “Nothing comes from nowhere.” This is true of PAD. The current PAD approach grew from a number of influences, some obviously traced but others more obscure. In this section, we will review the following influences on the development of PAD: the integration of psychology and psychometrics, the focus on evidential reasoning, and the influence of design science.

Integration of Psychology and Psychometrics

The writers and researchers advocating close integration of psychology and psychometrics have been an important influence on the development of PAD. Leveraging contemporary research on how students learn and build knowledge to define constructs distinguishes PAD from conventional approaches to test design. Early advocates often approached the integration of psychology and psychometrics from the perspective of validity (Anastasi, 1967, 1986; Lindquist, 1951; Loevinger, 1957; see Snow and Lohman, 1989, and Lawrence and Shea, 2008, for more extensive reviews). For example, Loevinger (1957) made the pithy observation about test development driven by criterion-related validity (here, *criterion* is used to mean what the test predicts well):

The argument against classical criterion-related psychometrics is thus twofold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful. (p. 82)

The 1980s saw a proliferation of authors both urging and demonstrating a closer coordination of psychology and psychometrics. An impetus for this movement may have been that psychology was escaping from the restrictions of behaviorism (Lachman et al., 1979). As Glaser (1981) noted, learning scientists had begun studying learning and individual differences in domains of interest to educators such as mathematics and reading, whereas past psychological research concentrated on the discovery of domain-agnostic mechanisms of thinking and learning. By the end of the decade, progress in the integration of learning science and psychometrics led Snow and Lohman (1989) to question the usefulness of conventional, trait-based, psychometric score interpretations:

The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts. Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs. (p. 317)

The 1990s saw the promise of integrating psychology and psychometrics turned into practical applications. Researchers proposing practical applications of an integrated approach included Lohman and Ippel (1993), who described a cognitive diagnostic framework for creating assessments that took advantage of research identifying item features that could be manipulated to vary cognitive complexity and item difficulty. Another researcher proposing an integrated approach was Nichols (1994), who claimed that theories of learning and cognition were well suited to informing assessment design

and development. Nichols proposed a framework for test development that consisted of an observation design and a measurement design. The observation design describes the characteristics of assessment activities, such as tasks or items, that make demands on the test taker, how these characteristics are to be organized in the construction and ordering of tasks or items, and the nature of the responses required. The measurement design defines the object of measurement and describes the procedures to assign a value or category to an object of measurement. In addition, Embretson (1998) described a cognitive design system consisting of a conceptual framework that integrated models of learning and cognition into test design and a procedural framework that described the steps necessary to implement the conceptual framework.

Focus on Evidential Reasoning

Another important influence on the development of PAD has been writers, particularly Schum (1994, 2009) and Toulmin (1958, 2006), who examined the evidence we gather and use as a basis for making claims in the fields of law and philosophy and that have been applied to assessment design beginning with ECD.

From Schum (1994) came an emphasis on first principles that encouraged PAD practitioners to step outside the implicit assumptions and customary practices of conventional psychometrics and reconsider the evidentiary reasoning underlying probabilistic inferences about emerging complex constructs. When reconsidering conventional psychometrics, PAD practitioners have borrowed from Schum the principles of relevance, credibility, and force of evidence. First, questioning the relevance of evidence has motivated practitioners to critically examine the bearing of evidence on intended score interpretation and use leading to, primarily, a greater inclusion of cognitive process evidence and less reliance on correlational and predictive evidence. Second, questioning the credibility of evidence has resulted in practitioners articulating and carefully examining the trait-based assumptions implicit in conventional psychometric practices. Finally, a skeptical stance toward the relevance and credibility of conventional psychometric evidence has led practitioners to reconsider the weight given to particular evidence in a score interpretation and use argument made within a principled assessment framework.

From Toulmin (1958, 2006) came an emphasis on practical arguments, such as those used in nonmathematical fields like law, rather than formal logic. When arguments are evaluated in formal logic, the goal is to determine whether an argument is true or false. Practical arguments cannot be usefully evaluated by the rules of formal logic because the assumptions used in practical arguments cannot be taken for granted; the available evidence is often incomplete or questionable. As such, the practical argument is, at best, convincing or plausible rather than true or false.

The combined influence of Schum (1994) and Toulmin (1958, 2006) contributed to the role of evidentiary reasoning as an integral part of PAD. The emphasis on making evidential reasoning visible through the explicit articulation of all assumptions, design decisions, and rationales for those decisions is a distinguishing characteristic of PAD.

Similar influences are apparent in the development of an argument-based approach to validity (Kane, 2006, 2013; Messick, 1989) and design propositions in design science (van Aken & Romme, 2009), creating compatibility with PAD.

Design Science

The emergence of design science as a field has been an important, if difficult to trace, influence on the development of PAD. In the seminal book *The Sciences of the Artificial*, Herbert Simon (1969, p. 113) argued for a science of design that is “tough, analytic, partly formalizable, partly empirical, teachable doctrine.” The response to this call was design science. The purpose of design science is to produce and communicate evidence-based design propositions (Johannesson & Perjons, 2014) of the form, “If you want to achieve Y in situation Z, then apply intervention X.” There are four defining characteristics of design science (Johannesson & Perjons, 2014) that influenced PAD. First, the goal of design science is to generate and test hypotheses about artifacts, such as assessment design patterns, as solutions to design challenges. Next, these hypotheses are tested and theory is built using rigorous research methods typically borrowed from the social sciences, including protocol analysis and field trials. Third, explicit prescriptions are offered on how to design and develop an intervention, such as game-based or three-dimensional science assessment. Finally, new findings on effective design are communicated to both practitioners (e.g., psychometricians) and researchers (e.g., learning scientists in areas such as learning progressions in science education).

This way of thinking in terms of design propositions supporting both the construction of assessments eliciting targeted knowledge and skills and the creation of evidentiary arguments supporting score interpretation is evident in PAD through the mindset of reasoning from imperfect evidence. In PAD, for example, design propositions undergird design patterns, and the latter are a link in the evidentiary argument supporting reasoning from imperfect evidence. These four defining characteristics of design science come into play not only in the design process, but also in the ongoing validation process, particularly with regard to the theory of action (addressed in a later section of this chapter).

PAD: ESSENTIAL ELEMENTS

To decide to engage in PAD does not mean that one has committed to a rigid set of processes in a particular order; PAD is not like putting together a piece of furniture from a kit with step-by-step directions. Rather, at its essence, it is “a mindset that results in a useful set of documentation” (M. Pearlman, personal communication, 2003). As described in the previous paragraph, that mindset is reasoning from imperfect evidence. And reasoning requires ruthless interrogation of ones’ assumptions, always coming back to the essential question, What is my rationale for this decision or action? This question can be asked of any action within PAD, for example, What research has informed this definition of the construct? Are these the knowledge and skills valued in

the discipline? Why did I choose to use this item type for this target of measurement? What evidence will this item yield to help me place the student along the underlying proficiency trajectory? Why do I believe that is so? And so on. At its heart, the PAD endeavor involves a team of interdisciplinary players who are committed to this type of thinking and discourse throughout the design and development process. When the answers to these questions are transformed into reference documents and design tools, and then the resulting documentation is used as the infrastructure for the inferential and validation arguments that are articulated and refined continuously throughout the design and development process, then one has engaged in PAD.

In what follows, we have distilled PAD to six essential elements organized into three general steps. It is critical to keep in mind that we have organized in this way for ease and clarity of description, but in practice the lines between the steps are blurred as the elements (or, to be more specific, the artifacts representing the elements) are refined through the iterative design and development process. These elements and steps represent the design and development process up to the first field-testing event. After data are collected, those data are used to inform revisions to the artifacts. This feedback loop of using data to inform refinement of the artifacts continues through the life of the assessment program.

The use of PAD will vary depending on the training, expertise, and philosophical orientation of the individuals comprising the assessment design team and the particular needs of the assessment program. For example, PAD will be implemented differently by a team who has extensive experience applying ECD or assessment engineering in a variety of large-scale assessment contexts versus by a team whose focus of applying PAD, ECD, or assessment engineering has been in a single domain (e.g., scientific inquiry in chemistry) and whose work has been designing rich, performance-based tasks for the classroom. The former may use terminology like “student model,” “evidence model,” “observable evidence,” “task model,” “difficulty drivers,” and “grammar,” whereas the latter may use terminology like “focal knowledge, skills, and abilities,” “unpacking,” and “design patterns.” These differences in terminology are simply different access points to the same desired outcomes of PAD, which we have distilled here into the six essential elements. Similarly, engaging in PAD will look very different when one is at the very beginning of designing an assessment versus whether one is using PAD to make improvements to an assessment program that has been operational for years. We posit that as long as some form of one or more of these essential elements is used—and the mindset is present—PAD will have some of the positive impact intended.

The three general steps are: (a) analyze the domain, (b) model the domain, and (c) create design patterns. As one moves through the steps, greater specificity and more attention to the constraints of the assessment are required. For example, in the domain analysis, one may articulate at a very high level the prioritized assessment targets (e.g., in mathematics, modeling, reasoning, and problem-solving are prioritized over procedural knowledge; algebraic reasoning is prioritized over geometric proofs). Then, as one moves into domain modeling, one must be much more specific to articulate the PLDs that will undergird the test and item specifications and support the intended score

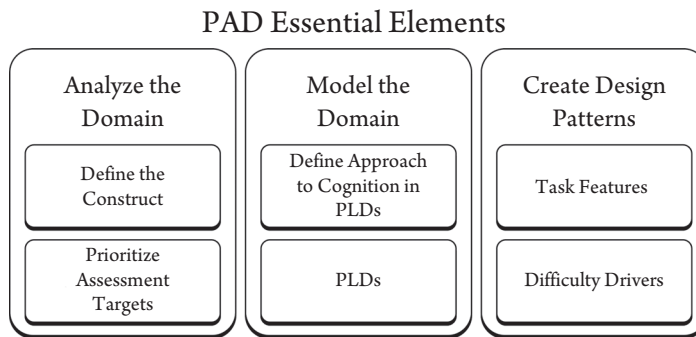


FIGURE 7.2
Principled Assessment Design Essential Elements

Note. PAD = Principled Assessment Design; PLD = Performance Level Descriptor.

inferences. Greater specificity still is required when creating design patterns because the constraints of the assessment must be considered when determining what kinds of tasks are best suited for the intended targets of measurement and how many tasks will be required to appropriately sample the domain and collect enough evidence to reliably support the intended claims about what students know and can do. Assessment design experts could reasonably disagree on when the assessment constraints should be considered in the design process. We argue that when an assessment is designed with the intention of cohering with instruction, the constraints of the assessment should be introduced at the domain model rather than during the domain analysis, and then the constraints are used more directly to shape the design pattern elements. In this way, the full breadth of the construct can be articulated and prioritized free of considerations of available item types and scoring mechanisms since those same constraints are not present in the classroom or in the discipline. Examples of richly articulated domain analyses can be found in the 2026 NAEP Mathematics and Reading Frameworks (NAGB, 2021a, 2021b), where the reader can readily envision reading and mathematics as sociocognitive and sociocultural endeavors in the classroom and in the real world without having to consider the respective constructs through a restrictive lens of constraints and features that are specific to the NAEP examination context.

See Figure 7.2 for the steps and elements of each step. Each will be discussed in detail in the following paragraphs.

ANALYZE THE DOMAIN

Define the Construct

It is unusual in the education context for the assessment developer to be charged with defining the construct, but when this is the case, the construct should be defined based on what we know about how students learn and build knowledge in the domain of interest and how that knowledge is used in the relevant discipline(s). In other words, for assessments of, say, biology, the construct should be defined based on what we know

about how students learn and build deep conceptual understanding in biology, as well as how this knowledge is used and employed in the biological sciences through the scientific practices. The role of learning science in the domain analysis cannot be understated for assessments that are intended to support inferences about what students know and can do, and even more so for assessments that are design to support instructional decisions (NRC, 2001; Perie & Huff, 2016). The theory of learning is the first segment of the assessment triangle from Pellegrino et al. (2011), cognition: What are our research-based hypotheses about how students learn and grow in the domain? It is this theory of learning that undergirds how cognitive complexity plays out in the domain model and, ultimately, the design patterns. For example, the NGSS are considered unprecedented in their complexity because they attempt to lay the groundwork for how science knowledge is not only acquired but also used in practice, which is why the standards are presented as performance expectations that reflect the integration of disciplinary core ideas (i.e., the learning that is required as the foundation for more learning), science and engineering practices (i.e., how knowledge is used and applied in the discipline), and crosscutting concepts (i.e., making connections across science domains is a way that students can make meaning of new information).

Another example of the essential role that learning science plays—or should play—in what we teach, learn, and assess has played out quite publicly between the mid-2010s and the mid-2020s in the United States (Goldstein, 2020; Hanford, 2018, 2019). As fourth-grade reading proficiency rates on NAEP continued to hover in the 30%–40% range for decades, Hanford (2018, para. 3) asked,

How do we know that a big part of the problem is how children are being taught? Because reading researchers have done studies in classrooms and clinics, and they've shown over and over that virtually all kids can learn to read—if they're taught with approaches that use what scientists have discovered about how the brain does the work of reading. But many teachers don't know this science.

Although a number of empirical studies have shown that children need explicit phonics instruction to learn to read, “there is a history of ignorance, complacency and resistance in colleges of education with regard to disseminating this critical information to pre-service teachers” (Hurford et al., 2016, abstract). This same kind of ignorance, complacency, and resistance is evident in conventional approaches to assessment when it comes to building on learning science research in our design endeavors. As this example demonstrates, it is critical that scientific research about how students learn in the domain of interest be the foundation of our work in the domain analysis, which paves the way for modeling the domain, where we articulate with more specificity the approach to cognition in the performance levels as well as how students move from novice to mastery with regard to each grade-level expectation.

Similarly, when defining the construct, we must consider that the assessment should yield valid interpretations for all students: students from various sociocultural backgrounds, multilanguage learners, and students with disabilities. Understanding from the

outset what the research indicates about how students from these various populations relate to learning the domain of interest is critical to defining the construct in ways that do not automatically include construct-irrelevant features for these students (Hong & Lissitz, 2017; Randall, 2021; Solano-Flores, 2019; see also Ercikan & Solano-Flores, this volume).

Assessment Targets

Once the construct is defined in this way, then the prioritized assessment targets can be defined at a high level. As mentioned in the section “PAD: Essential Elements,” we recommend that assessment constraints are not considered at this point, so that the assessment targets can serve not only as a basis for interim and summative assessment, but also for a variety of classroom and formative assessments. For example, in the redesign of the AP science courses, an approach was used to organize the domain into big ideas, enduring understandings for each big idea, and supporting understandings for each enduring understanding (Ewing et al., 2010). It was decided early in the design process that the target of learning, and therefore assessment targets, should be enduring understandings because these are the foundations that best prepare the learner for additional learning. Similarly, when the CCSS Mathematics were released, the guidance was for the major work of the grade, supporting standards, and additional standards to be proportioned 70%, 20%, and 10%, in curricular design, instruction design, and assessment design. These types of high-level guidance that are determined free of consideration of the interim or summative assessment constraints are critical to support coherence across instruction and assessment.

Most often in assessment of educational achievement, however, the analysis of the domain is provided to the test makers in the form of learning standards and test specifications that take the constraints and design features of the assessment into account. The constraints of the assessment program delimit the design features. Constraints and design features are determined by the answers to questions such as, but not limited to, the following.

The following are examples of constraints:

- How much time is there for design and development? What resources are available (e.g., personnel, funds)?
- What are the time limits for administration of the assessment? What is the expected turnaround time for score reporting?
- Must the assessment maintain a scale and/or comparability with an existing assessment, or is there freedom to depart?
- What item types are possible given the item authoring platform that will be used?
- What scoring mechanisms are available? For example, how much time and money exist for human raters, or must everything be machine scorable? What format(s) do the scoring mechanisms support (e.g., text, spoken word)?

The following are examples of design features:

- Is the assessment paper based or computer based?
- If computer based, is the assessment linear or adaptive?
- Does the client require a specified balance of item types?

We will return to the notion of constraints and design features in the section on “Design Patterns.”

When the assessment designer is provided the learning standards (either with or without test specifications) as a starting place, the PAD process can start with modeling the domain. If the learning standards were not developed from a research base of how students learn, then the relevant learning science research must be integrated in the domain model and design pattern elements. Sometimes this process is called *unpacking* (Harris et al., 2016, 2019).

When test specifications are also provided as a starting place, excavating the implicit judgments that undergirded the myriad decisions that were made to get from learning standards to test specifications is of utmost importance in PAD. Otherwise, the assumptions about the relationships among the learning standards and the evidence yielded by an assessment with the said specifications are left unarticulated, which undermines the coherence required for a strong inferential argument and a strong validity argument to support the resulting interpretations about what students know and can do.

Model the Domain

As stated previously, various instantiations of PAD throughout the years have taken different perspectives on what activities are associated with each step of the process and what artifacts are created in that step. For example, Mislevy and Riconscente (2005) used PAD as a tool for developing rich, computer-based performance tasks of scientific inquiry. In their approach, the domain model includes, among other elements, the assessment argument in Toulmin form (1958, 2006) and the design patterns. However, for the purposes of this chapter and in the context of designing interim and summative assessments of educational achievement for which one of the primary purposes is to classify students categorically along a latent performance trajectory, the two critical elements that must be articulated after the domain analysis is complete and before task design can begin are: (a) an articulation of the approach to cognition and (b) development of the PLDs.

Approach to Cognition

It is not yet common practice for educational assessment programs to identify explicitly the model of cognition that is implicit in the PLDs, the tasks, the evaluation of student work, the scoring model, and, ultimately, the interpretation of what students know and can do (Ferrara & DeMauro, 2006; Huff et al., 2017). This is true despite the great advances that cognitive research in how students process information has made since the publication of the fourth edition of *Educational Measurement* in 2006 and that of

the current volume in 2025. When we pay more attention to the role of cognition in learning, we unleash enormous potential to improve assessment design and sharpen our interpretations about what students know and can do when we pay more attention to the role of cognition (Huff et al., 2010; Leighton & Gierl, 2011; Mislevy, 2006; Nichols & Huff, 2017; NRC, 2001; Snow & Lohman, 1989). If the role of cognition is not explicitly engineered into the assessment design process, then we risk the emergence of an implicit, fragmented, and flawed proxy “model,” which most often leads to defaulting to the measurement of declarative or procedural knowledge or happenstance complexity (Huff et al., 2017).

We posit that given the constraints within which most educational assessments are designed and developed, it is likely to be a bridge too far to expect that a model of cognition be researched and defined *a priori*. However, it is essential that what we are calling the *approach to cognition* be articulated prior to (or in conjunction with) development of the PLDs. If not, then the assumptions about cognition will remain implicit in the individual minds of each player in the design endeavor, and it is unlikely that these implicit assumptions cohere. When models of cognition remain implicit in the minds of the various designers of PAD elements and consumers (primarily, educators), coherence is not achieved and the meaningfulness of the assessment results is compromised.

Before we describe what is meant by *approach to cognition*, a few comments on terminology are warranted. Table 7.1 gives a few illustrative examples of how related terms—*cognitive model*, *cognitive demand*, *cognitive complexity*, etc.—are defined by various authors. We recommend that for the purposes of most interim and summative

Table 7.1 Various Characterizations of Cognition Related to Assessment

Term	Citation	Definition
Cognitive model	Leighton & Gierl, 2007	A simplified description of human problem-solving of standardized educational tasks, which helps to characterize the knowledge and skills students at different stages of learning have acquired and to facilitate the explanation and prediction of students' performance
Theory of cognition	Huff et al., 2017	Prioritized value of knowledge and skills within a domain, organized to reflect structures and progressions along a learning trajectory, that articulates the precise knowledge and skills required to respond correctly to an assessment item
Cognitive demand	Perie & Huff, 2016	The degree to which tasks require more complex knowledge and skills for students to respond correctly and comprehensively
Cognitive complexity	Ferrara et al., 2014; Koedinger et al., 2012	The condition encoding requirements and the response requirements of the task; how many response operations are needed; interactions among different sources of complexity
Intrinsic cognitive load	Gillmor et al., 2015	Construct-relevant item features that contribute to item difficulty

educational assessments, becoming facile with these terms and definitions would be helpful but is not a requirement to achieve the baseline clarity and precision required for coherence. What is critical is that there is consensus on the assumptions undergirding student cognition for the domain of interest, that these assumptions are based in scientific research, and that the assumptions be explicitly reflected in the PLDs and design patterns. This degree of thought, discussion, and consensus among the design team regarding articulation of the approach to cognition reaps dividends as the team starts to author and review items from the perspective of whether the item yields sufficient evidence for the intended target of measurement, item clarity and quality, accessibility, sensitivity, bias, and cultural and linguistic responsiveness. When the PLDs and design patterns are authored with clarity on how student cognition is represented in the targets of measurement, conversations about construct-relevant versus construct-irrelevant variance start from a place of shared understanding rather than implicit assumptions.

There are at least three essential issues for the PAD team to debate and on which to reach consensus when building a shared understanding of the approach to cognition. The first is related to the fact that students can use many different cognitive processes and strategies to solve problems. For the intended purpose and use of this assessment, and given the prioritized knowledge and skills of the domain and the assessment targets that were defined in the domain analysis, what process and/or strategies will be assessed, if any? Or is it out of scope for the intended purpose and use of the assessment to prioritize process and/or strategies as targets of measurement? The intended purpose and use of the assessment will determine whether the process or strategy used to respond correctly to an item is also a measurement target. For example, in multiplication of multidigit integers, students may employ one of many strategies to correctly solve 125×10 . One common strategy is to regroup, for example, $(100 \times 10) + (25 \times 10)$. Another common strategy is to factor, for example, $125 \times 2 \times 5$. Whether strategies are a target of measurement must be articulated a priori because this decision will have an impact on the cognitive approach, PLDs, and design patterns. It is worth noting that if it is determined that processes and strategies are not prioritized as targets of measurement, assessment designers will still need to articulate and discuss the processes and strategies used by students because they are directly related to preconceptions and misconceptions, which need to be articulated as part of the design patterns.

The second issue for the team to discuss is the verbs that will be used to represent skills and how those verbs are defined in terms of observable evidence. The importance of using clear explicit action verbs cannot be understated (Egan et al., 2012; Perie, 2006; Perie & Huff, 2016). For example, verbs such as “know” or “understand” are ambiguous and are not nearly as readily interpretable as “identify,” “describe,” and “explain.” Another layer of clarifying verbs is to ensure that any synonyms are identified, for example, do we mean the same thing when we say *identify* versus *determine*? *Interpret* versus *analyze*? Perhaps one of the most critical elements of defining the approach to cognition is to define each verb in terms of observable evidence. In other words, if I asked a student to *interpret* a graph, what would I be looking for in the resulting interpretation

as evidence that the student interpreted rather than just, say, *described* the graph? What would I circle, underline, or highlight as evidence of an *interpretation*?

It is crucial that this thought process and discussion among the design team members occur without consideration of the assessment constraints, for two reasons. First, the approach to cognition is the first essential element of the domain model as it will directly inform PLDs. And, as cautioned earlier about introducing assessment constraints too soon, if the PLDs are to serve both curricular design and assessment design, and have reasonable generalizability to the domain of interest, they need to be written with applicability in the classroom as well as for the assessment (which does need to be addressed during PLD development, but not during approach to cognition; see the next section).

Second, suppose the assessment must be limited to two item types: multiple choice and short fill in the blank. If the team discussion about observable evidence of *interpretation* assumes the constraints of a particular item type, then the discussion is actually about what can be done within those item-type constraints. Establishing the key observable elements of the skill *interpret* should lead to the design of the item type or the selection of the best item type available to elicit the observable evidence of the skill. An example of this will be provided in the section on “Design Patterns.”

The third issue for the design team to discuss is: What assumptions do we have about how knowledge and skills relate to each other and cognitive demand? For example, in the assessment of historical thinking (as opposed to the assessment of recalling historical facts), it is generally agreed that the historical thinking skill “*identify and explain* historical developments” has less cognitive demand than “*analyze* patterns and connections between and among historical developments” regardless of the historical content, for example, pre-Columbian migration patterns or Victorian morality (Ercikan & Seixas, 2015), whereas in the sciences, the pairing of skill and content matters (Ewing et al., 2010). For example, the cognitive demand of accurately *describing* Heisenberg’s uncertainty principle is likely higher than the cognitive demand of, say, *analyzing* the equation $\text{force} = \text{mass} \times \text{acceleration}$, even though one would normally assume that *analyzing* requires more cognitive demand than *describing*.

Using skill taxonomies such as Webb’s Depth of Knowledge (DOK), Bloom’s taxonomy, SOLO taxonomy (Anderson & Krathwohl, 2001; Biggs & Collis, 1982; Webb, 2005), or others can be useful in considering the cognitive approach but are not sufficient, because the cognitive approach must be defined in light of the domain analysis—that is, in light of the prioritized knowledge and skills and in response to how students learn in the domain of interest. Given that most assessments of educational achievement are based on a set of learning standards that are not developed using PAD, it is likely that the design team will need to create an approach to cognition that is complementary to the learning standards. For example, suppose that the learning standards use a preponderance of ambiguous or less challenging verbs to represent skills (e.g., know, understand, identify, determine). The design team will have an opportunity to probe for clarity and desired level of rigor with stakeholders during the PLD development

process. Clarity will also be required when the learning standards include skills that are not generally fully assessed via typical exam constraints (e.g., evaluate, design, explore, create). The designers and stakeholders will need to reach agreement on the acceptable and reasonable evidence of these learning standards given the exam constraints.

Building consensus for—and documenting—the cognitive approach before developing the PLDs is critical because any latent ambiguity will cascade to the PLDs, the design patterns, and, ultimately, inferences about students. For assessment designers working on an assessment program that is already operational, where PAD was not present in the original design, all is not lost. There are at least two ways to infuse research-based approaches to cognition into an operational assessment program. First, most assessment programs need some new items each year. These consensus-building discussions that reveal latent assumptions about skills, verbs, and the relationships among skills and content can be used to improve item development specifications and item review criteria. Second, most educational assessment programs designed to classify students conduct standard setting on a specified cadence of every 3–5 years, or when needed, and most assessment programs take this as an opportunity to update PLDs (see Ferrara et al., this volume, for a discussion of standard setting). Updating PLDs is the perfect opportunity to infuse a transparent approach to cognition into the assessment design.

PLDs

PLDs are the claims about what students know and can do as they grow in mastery in the domain. In PAD, PLDs must be developed in advance of design patterns because they are essential elements in item design, scale development, and standard setting (Bejar, 2010; Bejar et al., 2007; Egan et al., 2012; Huff et al., 2017; Schneider et al., 2013). Generally speaking, PLDs reflect how knowledge and skills become more sophisticated as the achievement levels progress from representing *Novice* or *Emergent Mastery* through *Advanced Mastery*. The PLDs should emerge directly from the domain analysis and the approach to cognition that have been previously determined in the PAD process, both of which must emerge from the science of learning in the domain. Without this coherence, we risk that the inferences about where students are along the learning trajectory will not be useful to educators (Bejar, 2010; Frederiksen & Collins, 1998; Nichols et al., 2009; Schneider, 2017).

Egan et al. (2012) distinguished between the following PLD types and their uses in test development and score reporting: policy PLDs, range PLDs (RPLDs), target PLDs, and reporting PLDs. We discuss each in turn. More emphasis is given to the RPLDs given their role in item design, standard setting, and scale development.

Policy PLDs

Policy PLDs are high-level descriptions of what students should know and be able to do at each performance level. This description is developed for each domain and is typically common across grades. The policy-level definitions from NAEP Grade 4 Mathematics are shown in Figure 7.3. Policy-level PLDs are critical for setting the baseline expectations of the more detailed range, target, and reporting PLDs.

Grade 4

NAEP Basic (214)	Fourth-grade students performing at the NAEP Basic level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas. Fourth-graders performing at the <i>NAEP Basic</i> level should be able to estimate and use basic facts to perform simple computations with whole numbers, show some understanding of fractions and decimals, and solve some simple real-world problems in all NAEP content areas. Students at this level should be able to use—though not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses will often be minimal and presented without supporting information.
NAEP Proficient (249)	Fourth-grade students performing at the NAEP Proficient level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas. Fourth-graders performing at the <i>NAEP Proficient</i> level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the <i>NAEP Proficient</i> level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.
NAEP Advanced (282)	Fourth-grade students performing at the NAEP Advanced level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas. Fourth-graders performing at the <i>NAEP Advanced</i> level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

FIGURE 7.3

National Assessment of Educational Progress Grade 4 Mathematics Policy-Level Definitions

Note. NAEP = National Assessment of Educational Progress.

RPLDs

RPLDs are detailed descriptions of what students should know and be able to do in each performance level. RPLDs acknowledge that students within a given performance level represent a range of performance; for example, students at the low end of *Proficiency* have a different level of knowledge and skill than students at the high-end range of that level. Several factors should be considered in the organization and format of the RPLDs. First, the definition of the construct and prioritized assessment targets (the two essential elements of the domain analysis) and the approach to cognition (the first essential element of the domain model) should inform the structure and grain size of the RPLDs. Both the Partnership for the Assessment of Readiness for College and Career (PARCC) and the New York State Testing Program (NYSTP) used different elements of PAD in their development (Huff et al., 2017). In each case, the testing programs were provided with the domain analysis in the form of the CCSS that were designed to inform both curriculum and assessment and contained some high-level guidance on which standards should take priority. The CCSS also provided a starting

PARCC PLD Example

Grade 3 Math: Content (Sub-Claim A) The student solves problems involving the Major Content for the grade/course with connections to the Standards for Mathematical Practice.			
Exceeds Expectations	Meets Expectations	Approaches Expectations	Partially or Does Not yet Meet Expectations
Products and Quotients: 3.OA.1, 3.OA.2, 3.OA.4, 3.OA.6, 3.OA.7-1, 3.OA.7-2			
Understands and interprets products and quotients of whole numbers.	Interprets products and quotients of whole numbers.	Interprets products and quotients of whole numbers.	Determines products and quotients of whole numbers within 100.
Determines the unknown whole number in a multiplication or division problem by relating multiplication and division. Both factors are greater than 5 and less than or equal to 10.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division. One factor is greater than or equal to 5.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division, with both factors less than or equal to 5, or with one factor of 10.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division, with both factors less than or equal to 5, or with one factor of 10.
Represents a multiplication or division situation as an equation.			
Accurately multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	Accurately multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	Multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	
Multiplication and Division: 3.OA.3-1, 3.OA.3-2, 3.OA.3-3, 3.OA.3-4			
Uses multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Both factors are greater than 5 and less than or equal to 10.	Uses multiplication and division within 100 to solve word problems involving equal groups and arrays. One factor is greater than or equal to 5.	Given a visual aid, uses multiplication and division within 100 to solve word problems involving equal groups and arrays, with both factors less than or equal to 5, or with one factor of 10.	Given a visual aid, uses multiplication and division within 100 to solve word problems involving equal groups. Both factors are less than or equal to 5, with both factors less than or equal to 5, or with one factor of 10.

NYSTEP PLD Example

Grade 3 Mathematics Performance Level Descriptions				
Cluster	Performance Level 4	Performance Level 3	Performance Level 2	Performance Level 1
Students represent and solve problems involving multiplication and division. (3.OA.1-4)	Interpret and represent products and quotients of whole numbers. Determine the unknown whole number in a multiplication and division problem by relating multiplication and division. Represent a multiplication or division situation as an equation. Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Identify proper context given a numerical expression involving multiplication and division. Both factors are less than or equal to 10.	Interpret products and quotients of whole numbers. Determine the unknown whole number in a multiplication or division equation relating three whole numbers by relating multiplication and division. Factors are greater than 5 and less than 10. Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Both factors are less than or equal to 10.	Interpret products of whole numbers. Determine the unknown whole number in a multiplication equation by relating multiplication and division. Limit to factors less than or equal to 5. Given visual models and/or manipulatives, use multiplication and division within 100 to solve word problems involving equal groups and arrays. Both factors are less than or equal to 10.	Given visual models and/or manipulatives, interpret products of whole numbers with factors less than or equal to 5. Determine the product in a multiplication equation with whole number factors less than or equal to 5. Given visual models and/or manipulatives, compute products within 25 in the context of word problems.
Students understand properties of multiplication and the relationship between multiplication and division. (3.OA.5, 6)	Justify the use of properties of operations (commutative, associative, and distributive) as strategies to multiply. Restate a division problem as an unknown factor problem, and explain the relationship between division and finding an unknown factor.	Apply properties of operations (commutative, associative, and distributive) as strategies to multiply. Restate a division problem as an unknown factor problem.	Apply the commutative property as a strategy to multiply. Restate a division problem as an unknown factor problem.	Given visual models and/or manipulatives, identify equivalent expressions that illustrate the commutative property within 10.

FIGURE 7.4**Range Performance-Level Descriptor Examples**

place for the approach to cognition in the form of the mathematical practices and an articulation of the relationship among the reading standards and text complexity. Figure 7.4 shows the different ways that the RPLDs were organized by PARCC and NYSTP; different approaches were used, but each was reasonable. For PARCC, the performance levels were organized by standard, whereas for NYSTP, cluster was the organizing unit. NYSTP also provided a bit more granularity in the PLDs than PARCC.

It is important to consider whether the RPLDs will be used to support curriculum, instruction, and broader assessment uses in the classroom or just the interim or summative assessment that will need to meet large-scale assessment psychometric and validation requirements. In a coherent educational system, the promise is that the RPLDs will be used to support both the classroom and the large-scale assessment, and if that promise is to be realized, then issues related to the verbs used to represent skills and the difference between assessment constraints and classroom constraints need to be considered in advance of RPLD development.

For example, take the two skills “identify” and “name” into consideration with regard to different kinds of parallelograms. Let’s suppose identify is defined in the approach to cognition as a student’s ability to select the correct term from a set of options (e.g., square, rectangle, rhombus) when presented with a shape with particular features, whereas name is defined as the student’s ability to independently generate the correct term. Unless ground rules are laid in advance, the RPLD developers could introduce unnecessary limits on the RPLDs by taking the assessment constraints into account. That is, if it is known in advance that the assessment will not be able to accommodate constructed response or speech recognition for name, then some would argue that because the RPLDs will serve item design, standard setting, and scale development, then the skill “name” should not be used in the RPLDs. However, that would make the RPLDs less useful for the classroom, which does not have the same constraints as the assessment setting. As such, we recommend that when the intent is for the RPLDs to serve both the classroom and the large-scale assessment, the RPLDs be written without the assessment constraints taken into account. We argue that the resolutions required for the RPLDs to serve item development, standard setting, and scale development can happen at a later time. For standard setting, a skills glossary can be provided for panelists that indicates, for example, “for the purposes of this methodology, assume the verb *identify* each time the verb *name* is used,” or the verbs can be changed accordingly in the PLDs that will be used in standard setting. For item and scale development, the design pattern is where resolutions between targets of measurement (skills, knowledge) and assessment constraints (e.g., item types) are addressed.

Claims about students in the RPLDs should reflect the approach to cognition previously articulated, reflect the knowledge or content prioritized in the domain analysis, and identify any relevant context that has an impact on the cognitive demand, and therefore the performance level in which the student claim is situated (Egan et al., 2012; Ferrara & Steedle, 2015; Ferrara et al., 2015; Schneider et al., 2013; Valencia et al., 2014). Before giving an example, it is important to note that the development of

the RPLDs will likely require some iteration with the approach to cognition. It is still advisable to document the approach to cognition prior to RPLD development so that all are working from a shared understanding and a shared vocabulary, which will facilitate productive iteration as opposed to unproductive churn.

Table 7.2 provides an example of RPLDs where the approach to cognition, knowledge (or content), and context is embedded in ways that are intended to model a progression of cognitive demand, from early grade-level mastery (left column) to advanced grade-level mastery (right column).

Typically, students in the earlier stages of mastery need reduced cognitive load to demonstrate what they know and can do with grade-level content and skills, so it is critical to include this kind of context at the appropriate RPLD performance level to differentiate on grade-level performance with scaffolds from below-grade-level work.

Target PLDs

Rather than the range of performance within a particular level, the target PLDs (sometimes referred to as the threshold PLDs) focus on the distinguishing characteristics

Table 7.2 Example Range Performance-Level Descriptors

Approach to Cognition: Verb Used to Represent the Skill		
Knowledge/content: The similarities and difference among similes, metaphors, allusions, alliteration		
Context: Text (stimuli) can require different levels of inference		
Early grade-level mastery	Mid-grade-level mastery	Late grade-level mastery
Student identifies an author's use of easily inferred figurative language or other literary devices (e.g., similes and metaphors) in a literary text.	Student identifies an author's user of figurative language or other literary devices (e.g., allusions and alliteration) and interprets how this language contributes to the meaning of a literary text (e.g., how a metaphor expresses a theme).	Student identifies an author's user of figurative language or other literary devices and evaluates how this language contributes to the meaning of a literary text that is complex and requires a high degree of inference.

at the “threshold” of each performance level. This facilitates necessary discussions in various standard-setting procedures where panelists need to reference a shared understanding of what knowledge and skills represent a student who is “just barely” in one level as opposed to the one just preceding it.

Reporting PLDs

Once the cut scores are finalized, any adjustments that need to be made for reporting what students in each performance level know and can do should be made. For example, a decision could be: Should the reporting PLDs reflect the minimally sufficient knowledge and skills of the placement level or the typical knowledge and skills of the placement level?

DESIGN PATTERNS

The third step in PAD—creating design patterns—has two essential elements: (a) an articulation of the task³ features that are required to elicit the evidence necessary to support the intended claim about the student as articulated in the RPLDs and (b) an identification of the task features that impact cognitive demand (frequently referred to as difficulty drivers) and, by extension, item psychometric characteristics and scale properties. Both of these essential elements build on essential elements that have been articulated previously in the PAD process: the definition of the construct, the prioritized assessment targets, the approach to cognition, and the RPLDs. There are many other elements of the assessment design endeavor that will need to occur after RPLDs are authored before an assessment is ready for field testing; Figure 7.5 provides a summary.

We have chosen to highlight task features and difficulty drivers as essential elements because they are the unique features of PAD that typically remain implicit in conventional approaches to assessment design. As with other aspects of PAD, the

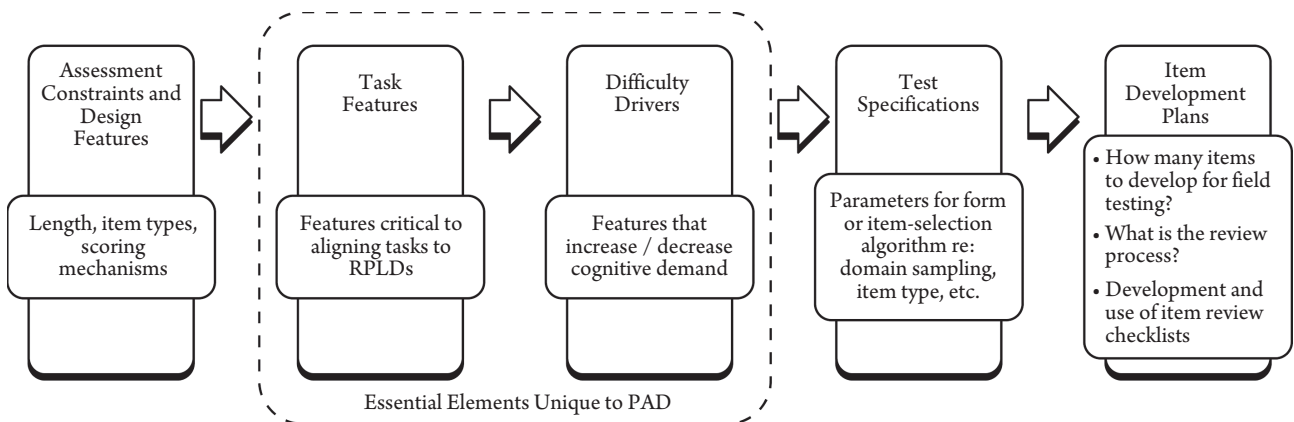


FIGURE 7.5
Remaining Steps in Assessment Design

Note. PAD = Principled Assessment Design; RPLD = range performance level descriptor.

terminology for the process of task design used in various publications and applications can be daunting and off-putting (e.g., design patterns; task models; task features; task specifications; variable features; fixed features; focal knowledge, skills, and abilities; assessment framework). In an attempt to clarify, we posit that some of the key benefits of PAD can be achieved if the task features and difficulty drivers are carefully articulated, used to develop items, and refined accordingly as data are analyzed as part of iterative, ongoing design. For operational testing programs not developed under PAD, infusing task features and difficulty drivers into the item design specifications can be done over time.

One of the most challenging aspects of creating design patterns is what grain size to start with: A single learning standard? An RPLD component? A single target of measurement for a single item? As Hendrickson et al. (2010) described, these decisions are arbitrary and the process is iterative. A good starting place will be influenced by the structure of the RPLDs and the assessment constraints and design features. For example, the NYSTP Grade 3 Mathematics RPLDs are organized by cluster (an organizing feature within CCSS) and then become more granular as the progression is articulated against four performance levels (see Figure 7.4). The first row of the first cluster has to do with *products and quotients of whole numbers* and is ordered from below grade-level performance to advanced grade-level performance as such:

- Level 4: Interpret and represent products and quotients of whole numbers.
- Level 3: Interpret products and quotients of whole numbers.

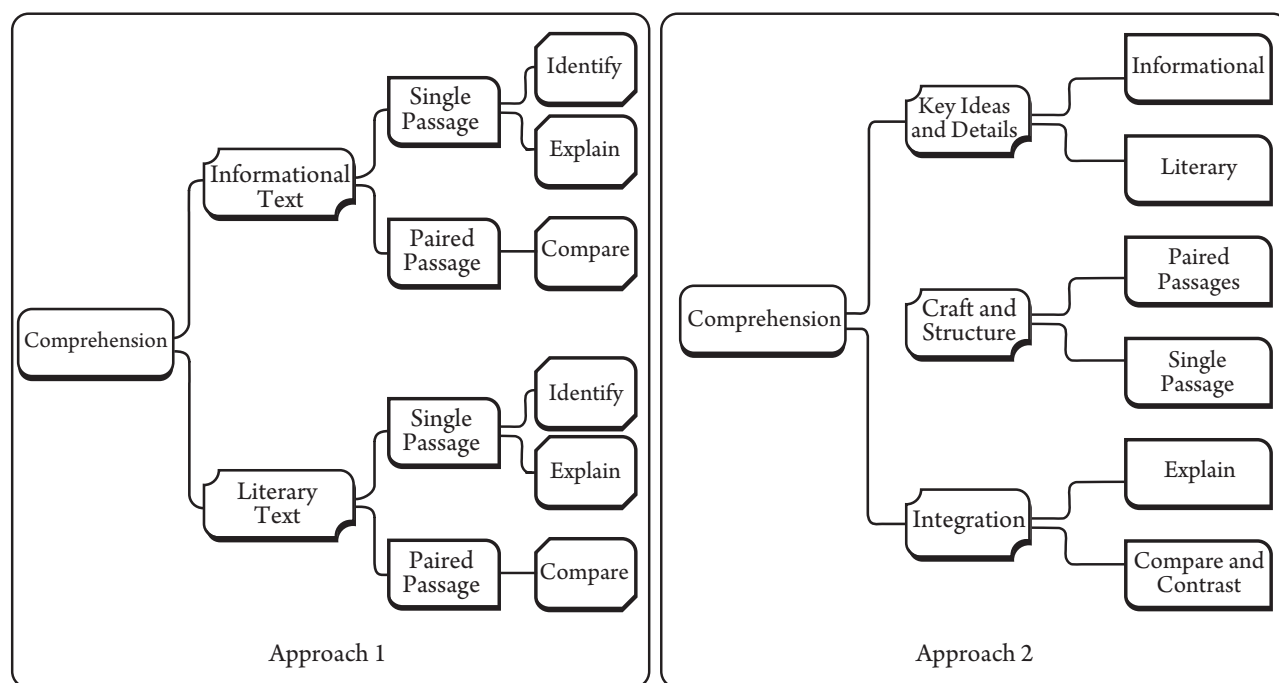


FIGURE 7.6

Design Pattern Organization and Grain Size

- Level 2: Interpret products of whole numbers.
- Level 1: Given visual models and/or manipulatives, interpret products of whole numbers with factors less than or equal to 5.

It seems that the first row of the first cluster would be a likely place to start and yield a useful grain size for a design pattern. However, if time and resource constraints are a factor, a design pattern that applies to the full cluster will be sufficient and better than no design pattern at all. Figure 7.6 provides an example of two different approaches to creating design patterns for a reading comprehension domain. In the first approach, the design patterns are organized by text genre, number of passages in the stimuli, and then specific skill. In the second approach, design patterns are organized by skill clusters and then by other features depending on the cluster. The take-home message from this example is twofold: There is variability and choice in the grain size of design patterns depending on the specific needs of the testing program and the resources available; and both the organization and the grain size of the design patterns are arbitrary. What is essential is that the task features that anchor items to the RPLDs and the difficulty drivers that impact cognitive demand are articulated prior to beginning item development. Although task features and difficulty drivers may change depending on whether one is working with a single informational passage where the target of measurement is identify versus explain, if there are not resources to create design patterns at that level of specificity, then having a design pattern one or two layers up is still beneficial. We must start where we can and claim small PAD victories along the way.

Task Features

Articulation of the task features provides the essential, explicit link between the RPLDs and the tasks to be written by explicating the features of stimuli (e.g., text, graph, video, map, diagram), prompt (the question or problem to be solved), key (the correct answer), distractors (incorrect answers based on preconceptions and misconceptions), and other ways in which unique aspects of the item type can be manipulated to help ensure that each item captures evidence relevant to placing the student along the latent performance continuum (Hendrickson et al., 2010; Luecht, 2013, 2019). Each feature of the task contributes to the cognitive demand of the task (Ferrara et al., 2015; Koedinger et al.; 2012; Schneider & Johnson, 2019), so each must be considered carefully in relation to the RPLDs. The constraints of the testing program will delimit each of the task features. For example, an assessment program that is restricted to machine-only real-time scoring and must be relatively short in duration (e.g., an adaptive interim assessment) will have task features that differ from an assessment program—of the same content and skills—that includes, say, essays or computer-based simulations. Each decision about task features should be warranted by the learning science research that undergirds the domain analysis and the domain model.

For example, a design pattern for the RPLDs shown in Table 7.2 would need to address, at minimum, the task features that differentiate performance between the levels with greater specificity. Currently, the RPLDs suggest that similes and metaphors

are associated with a different RPLD than allusions and alliterations. Given that similes, metaphors, allusions, and alliterations are provided as examples, what about other figurative language devices such as personification or hyperbole? Is each on grade level for this standard, and if so, are they associated with a particular RPLD? Other features that should be addressed in the design pattern include the features of text that must be present to support the kinds of questions that students will be asked to respond to in relation to the text, such as, What are features of text that make figurative language “easily inferred” as opposed to requiring a “high degree of inference”?

Also to be addressed in the design pattern are how different item types that are available for the assessment relate to the intended targets of measurement and performance levels. For example, the assessment design team may determine that the skill identify can be assessed with fidelity with either multiple choice or drag and drop, whereas the skill evaluate requires students to highlight relevant text in the passage. In each case, the rationale for the use of the item type(s) for the skill should be documented in the design pattern, as well as any specific features of the item type that must be present to elicit evidence for that particular skill. In addition, we need to refer to the approach to cognition when creating design patterns. In the approach to cognition, we define the observable features of the skills, in this case what constitutes an *identification* and an *evaluation*. One common characteristic of evaluation that distinguishes it from identification is a judgment of quality. In the design pattern, we articulate how the task features elicit from the student evidence of the RPLD claim (or part of the claim) that is the target of measurement, in this case, evidence that gives one confidence that the student is *evaluating* rather than *identifying*.

Difficulty Drivers

In the previous section, we asserted that an essential element of design patterns is to articulate the task features that elicit evidence that a student is in one placement level versus another. When the RPLDs are constructed carefully, the distance to get from RPLDs to design patterns is a few steps rather than a journey. A second essential element of design patterns is difficulty drivers, that is, What are the task features that we can manipulate to alter cognitive demand but stay within the intended RPLD? Recall that any performance level represents a range of proficiency, and items are needed that represent the range. Identifying the task features that can be manipulated to adjust cognitive demand is helpful to support development of items that will span a range of observed difficulty. When considering the relationships among task features and cognitive demand, it is imperative to consider the diversity of the student population such that we are avoiding construct-irrelevant difficulty for students from various sociocultural backgrounds, multilanguage learners, or students with disabilities.

Figure 7.7 shows an excerpt from a detailed design pattern from an AP Biology research study on automatic item generation (Huff et al., 2013). Here, the researchers identified six ways to manipulate the difficulty of items while still targeting the claim within the *Basic* performance level. Also noteworthy is that the researchers identified two task features that do not impact cognitive demand, which is also helpful information. Note that evidence

Construct Identifier:	Biology
Primary Context:	Cell division
Competency Claim	The student can construct explanations of how DNA is transmitted to the next generation via the processes of mitosis, meiosis, and fertilization.
Proficiency Level	Basic
<i>Evidence Documentation</i>	
1.	Description of the purpose of mitosis and meiosis
2.	Description of the products of mitosis and meiosis
3.	Description of the behavior of the chromosomes during the phases of mitosis and meiosis
4.	Explanation of the processes of mitosis and meiosis
5.	Comparison and contrast of the processes of mitosis and meiosis
6.	Use and recognition of vocabulary specific to cell division
<i>Manipulable features of complexity</i>	
1.	Type of cell division (mitosis is simpler than meiosis)
2.	Number of steps in process (mitosis has fewer steps than meiosis)
3.	Type of statement/alternative (definition is less challenging than explanation)
4.	Use of vocabulary particular to cell division will increase complexity: ploidy, tetrads, synapsis, crossing over, sister chromatids, homologous chromosomes, segregations, equatorial plate, cytokinesis
5.	Phase of cell division in question; the events in some phases are more conceptually difficult than the events of other phases
6.	Making a comparison (more challenging) vs. selecting a true statement (less challenging)
<i>Features irrelevant to complexity</i>	
1.	Number of chromosomes in a cell
2.	Type of organism in which the processes occur

FIGURE 7.7

Snapshot of Design Pattern With Manipulable Features of Complexity (Difficulty Drivers)

documentation in Figure 7.7 is the observable evidence of an explanation within this particular context of cell division; this is another example of how different approaches to PAD have the essential elements associated with different processes. No matter the differences, the intent is the same: to make transparent all that is typically implicit in task design so that we can debate our assumptions, document our consensus, and use those documents as design tools and as evidence of our inferential chain of reasoning.

Ferrara et al. (2014) worked with a team of researchers across many testing organizations and practitioners across many states to develop an approach to considering cognitive complexity for the CCSS. The goals for these new metrics were to provide a systematic and replicable way of determining cognitive complexity for each task and to provide measurement precision at all levels of the score scale. Figures 7.8 and 7.9 provide taxonomies of cognitive complexity for English language arts (ELA)/literacy and math, respectively. Note that for ELA/literacy, the complexity and length of the text is considered a source of cognitive complexity along with the processing demands of the task (e.g., the

Proposed Sources of Cognitive Complexity for Items and Tasks: ELA/Literacy (Summary)

The goals and uses of cognitive complexity are:

- Provide a systematic, replicable method of determining item cognitive complexity
- Provide measurement precision at all levels of the test score scales

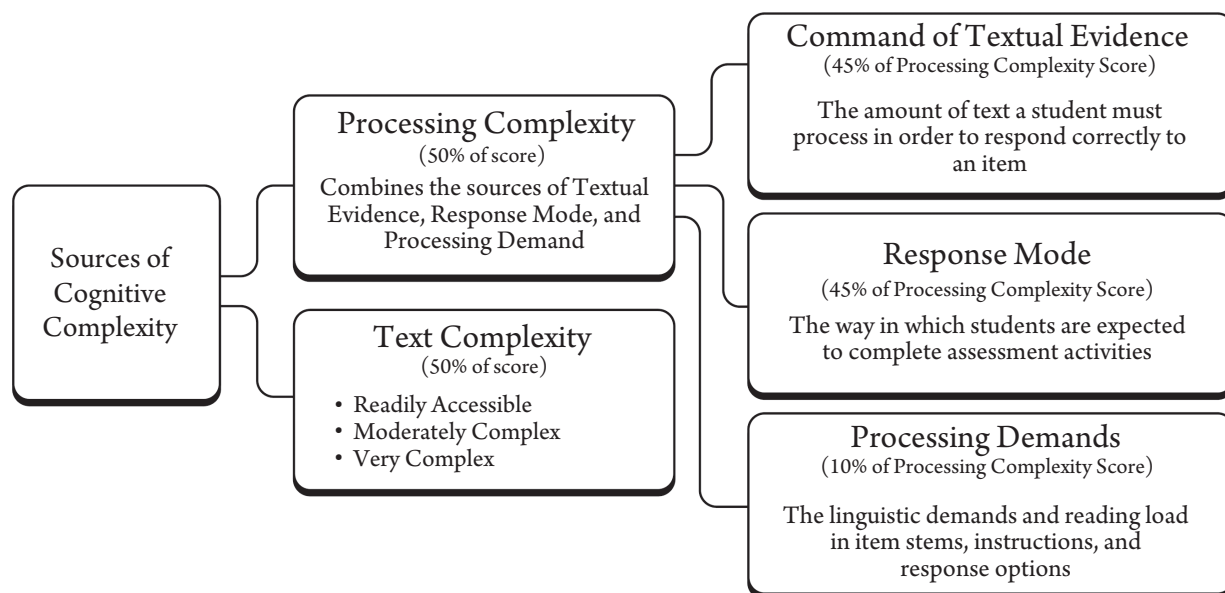


FIGURE 7.8

Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers English Language Arts/Literacy

stimulus and prompt presented to the student), as well as the response demands of the student (e.g., selected response versus constructed response). Similarly, for math, both stimuli demands and response demands are considered sources of complexity, as well as which mathematical practice is required in the task. For both ELA and math, there exist research-based rules for how to roll up these various features into a single metric of low, moderate, and high cognitive complexity that can be used in design patterns as a more sophisticated approach to difficulty drivers than conventional practice.

While this section has focused on design patterns, PAD does not negate the need to implement training to produce items of good technical quality and fairness. Readers desiring additional information in these areas should refer to Schmeiser and Welch (2006) or the *Handbook of Test Development* (Lane et al., 2016). The development of items must include focus on the accuracy, clarity, and accessibility of the items; the items being free of sensitivity and bias; and the items being culturally and linguistically responsive. We argue that when the PAD mindset of reasoning from imperfect evidence is shared by the design team and some of the PAD essential elements are developed and used as design tools, the resulting items have a strong pedigree—that is, a strong evidential argument regarding the validity of inferences we need to make about what students know and can do.

Proposed Sources of Cognitive Complexity for Items and Tasks: Mathematics (Summary)

The goals and uses of cognitive complexity are:

- Provide a systematic, replicable method of determining item cognitive complexity
- Provide measurement precision at all levels of the test score scales

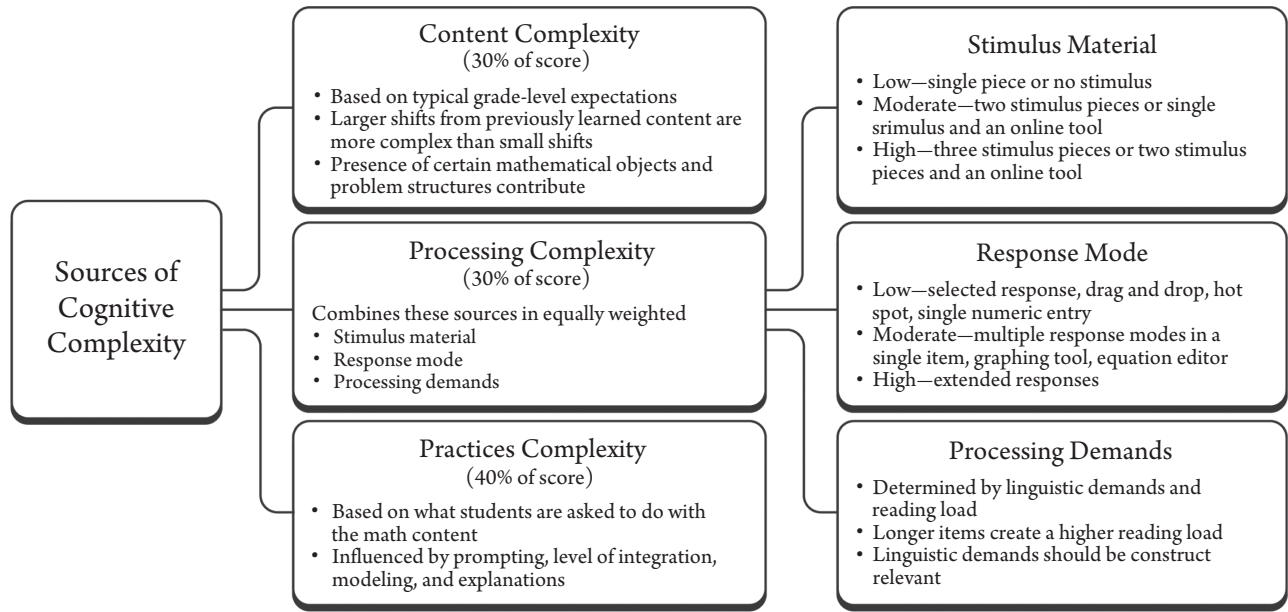


FIGURE 7.9

Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers Math

INTEGRATION OF DESIGN AND PSYCHOMETRICS

One of the primary advantages to using PAD is the integration of several aspects of test design that are distinct and even siloed in conventional approaches. Integration of skills and content into claims about students, integration of RPLDs into item design, and, finally, integration of item design and scale design are beneficial in multiple ways and represent a true evolution in the assessment industry.

Although standards-based assessments or, more specifically, assessments intended to support inferences regarding the degree to which students have met learning objectives have grown in demand and prominence from the 1980s to 2025, the publication date of this volume, the way we design and develop large-scale educational assessments is still largely situated within philosophies, structures, and processes that evolved to support assessments for which normative information was the primary inference of interest, such as percentile rank with regard to a specific population. Item development processes and psychometric research on these norm-referenced tests—SAT, ACT, Iowa Tests (formerly Iowa Tests of Basic Skills), TerraNova, SAT-10, and GRE, to name a few from the field of education—have informed much if not most of educational

measurement and educational test development. When testing companies—and the psychometricians and item writers⁴ in those companies—shifted from building “off the shelf”⁵ or selection tests to tests based on state-specific learning standards with the primary purpose of classifying students into proficiency categories, the required shifts in philosophy and practice were minimal at best.

The primary metrics of conventional norm-referenced K–12 tests to support inferences about students are percentile-rank based on a nationally representative population and grade-equivalent scores—another norm-based metric devoid of any connection to learning standards—not inferences about what students know and can do with regard to a set of learning standards that represent what is valued in the classroom and the discipline. For college or graduate school admission tests, the primary metric to support inferences is, again, percentile-rank based on a nationally representative population, and the primary validity evidence is prediction of future grades. From a business perspective, it only made sense to keep as many practices and processes in place as possible during this pivot into the new space of state-specific standards-based testing.

Many test designers and psychometricians, with the best intentions, likely assumed that the only necessary changes to traditional practices concerned producing what the states needed to meet federal requirements: a post hoc alignment of banked items from these off-the-shelf tests to the state’s learning standards and a standard setting after the first operational administration to establish cut scores for the performance levels. As discussed in the second paragraph of the chapter, the cry for a different philosophy and practice for educational assessment design gained strength and numbers during this era and culminated with the NRC’s *Knowing What Students Know* (NRC, 2001). At the time of this writing over 20 years later, practitioners are either still fighting for evolution in the direction of PAD for assessments designed to tell us about what students know and can do or maintaining status quo either out of a lack of knowledge of a different way or out of a belief that the cost of evolution in design practices is neither warranted nor worthwhile from a business perspective. The authors of this chapter are firmly in the first category.

As mentioned at the beginning of this chapter, the conventional approach to test development, if viewed from a high level, may not appear to differ from a PAD approach. For example, below is a synopsis of the main phases of the test development process as outlined by Schmeiser and Welch (2006):

Step 1: Identify intended purpose and use

Step 2: Determine test design features and constraints (e.g., time to test, time to score, paper delivery or computer based, item type balance)

Step 3: Test specifications

Step 4: Item development

Step 5: Field testing

Step 6: Evaluation of test results

The PAD approach is essentially the same as the conventional approach when viewed at this grain size, especially for Steps 1, 5, and 6. When one goes into the details of the processes and documentation of Steps 2–4, however, there are many differences, especially for educational achievement tests that require performance levels and standard setting. One key operational and pragmatic differentiator between conventional approaches and PAD that has a ripple effect throughout the design and development process—as well as implications for the inferential and validity arguments—is that the former has distinction, if not fragmentation, in areas whereas the latter has integration and coherence. Three related areas of integration that are hallmarks of the PAD approach are: (a) the integration of content and skills to be assessed in the test specifications; (b) the integration of PLDs into the item design process; and (c) using PAD to engineer desired scale properties. These types of integration require a deeper partnership among item writers and psychometricians—as well as representatives from other related disciplines who should also have a seat at the table—during the design phase of the assessment than is typically the case.

Data to Inform Design and Development

In Schmeiser and Welch (2006), field testing is noted as Step 5. Field testing (sometimes referred to as field trials, pretesting, or item tryouts) is generally defined as when items are administered to a sufficient number of students to support initial psychometric analyses. New or trial items do not count toward a student's score, primarily to avoid miscalculating the student's score should subsequent psychometric analyses reveal a flaw with the item or because the trial item is not yet properly scaled with all the other items on the test. The sample size for field testing will depend on the intended psychometric analyses; for example, classical item statistics do not require as many student responses as item response theory–estimated parameters. In PAD, just as in conventional test design, data from a well-designed field test are required to do the psychometric analyses described in this section. However, there are various opportunities for gathering data throughout the design process prior to field testing that should be taken.

For example, administering early prototypes of items to relatively small samples of students in settings where students can be observed and interviewed (e.g., cognitive labs) can confirm or help us revise our assumptions about the cognitive processes employed in responding, or the interaction between cognitive demand and item type. We can use this information to update our artifacts that represent the essential elements of PAD: approach to cognition, RPLDs, and/or our task models and difficulty drivers. For computerized assessments, it is strongly advised to frequently observe students navigating through multiple components of the assessment as they are developed, with a keen eye toward how the accessibility features are performing for students with various needs. When time and resources are available, using tools such as eye-tracking equipment can also provide fascinating insights into how students are engaging with the assessment and help improve how the tasks and navigation are designed. At the time of this writing, we are beginning to see testing programs gather input from students on item design as

part of an effort to better represent diversity and inclusion in our assessments. In each of these cases, the data from these opportunities serve as invaluable feedback to the design process and only serve to strengthen the evidentiary argument for the assessment.

Content and Skill Integration

In PAD, the content and skills to be assessed are integrated in the form of claims—for example, a student can classify two-dimensional figures in a hierarchy based on defining attributes, where “classify” is the skill and the content is “two-dimensional figures in a hierarchy based on defining attributes.” As discussed in the previous section, these claims are arranged along a progression, in the form of RPLDs, that demonstrate how a student moves within a school year from novice to grade-level proficiency. RPLDs in and of themselves cannot serve as test specifications because test specifications need to consider the number of items, the type of items, and other design features and constraints of the assessment to guide form assembly in a fixed-form assessment or the item selection algorithm in an adaptive testing environment. In a conventional approach, although content and skills may be weighted or balanced in the test specifications (e.g., the 12 items in the geometry domain should have a skills distribution of 30%–40% application and 60%–70% comprehension), the content and skills are not integrated as a claim about what students know and can do. As we argue in the section “Using RPLDs to Guide Assessment Design,” this lack of integration does not support the type of scale properties that are optimal for supporting performance standards (also referred to as cut scores) and inferences about what students know and can do that are based on the RPLDs.

Figure 7.10 provides two examples of typical mathematics test specifications for state accountability tests for elementary grades. In each case, the testing program goes on to give descriptions of each component in some detail but falls short of articulating how the content and skills (represented in each case by DOK) should be integrated to support the intended inferences regarding student proficiency. For instance, for the example on the right side of Figure 7.10, the technical manual includes descriptions of both DOK and the mathematical practices, and it indicates that two mathematical practices, *justification and explanation* and *modeling*, result in reportable scores but does not specify guidelines for how the mathematical practices and the DOK levels relate to one another or how they should be integrated into the assessment design. Without a priori hypotheses about how these various design components relate to eliciting the required evidence from students documented in design patterns or other item-writing specifications, these decisions are left to each individual item writer. Regardless of the depth of expertise of any given item writer, it is extremely unlikely that a coherent perspective on the following will occur by happenstance:

- Which item types and item features are best suited to elicit the types of skills implied by DOK 1 versus DOK 2 versus DOK 3?
- What is the best way to elicit evidence of mathematical practices given the item types and manipulatable item features available?
- What are the relationships between DOK and mathematical practices?

Example 1

DOMAIN	MIN.	MAX.
Operations and Algebraic Thinking	29%	38%
Number and Operations—Base 10	18%	22%
Number and Operations—Fractions	27%	31%
Measurement, Data, and Geometry	18%	22%
DOK 1	18%	31%
DOK 2	38%	58%
DOK 3	9%	20%

Example 2

OPERATIONAL ITEMS	NUMBER OF ITEMS
Item Types	
Multiple Choice	21–23
Technology Enhanced	0–6
Constructed Response	3
Reporting Categories	
Progress to Grade Level	19–20
Number and Ops—Base 10	3–5
Number and Ops—Fractions	4–5
Ops & Algebraic Thinking	3–6
Geometry	3–4
Measurement & Data	3–4
Mathematical Practices	
Integration	7–8
Justification/Explanation	3
Modeling	≥7

FIGURE 7.10**Conventional Test Specifications With Separate Considerations of Components**

Note. DOK = Depth of Knowledge.

- What is the optimal way to pair content and skills (or, in these examples, DOK level) given the intended inferences about student proficiency?

In their chapter on the role of cognitive models in large-scale educational test development, Huff et al. (2017) contended that with PAD,

all assumptions, especially those regarding student cognition, are brought to the foreground and examined. As a result, when the cognitive model is selected a priori and is made explicit, there is the opportunity to ensure alignment of the cognitive model with the intended purpose and use of the assessment as well as with all other design, scoring, reporting and interpretation decisions to yield evidentiary coherence. (p. 401)

In contrast, assessment programs that employ PAD in some way tend to have test specifications that integrate the content and skills to be measured in ways that make transparent the assessment designers' perspective on what is valued in the domain. The NAEP Technology and Engineering Literacy assessment specifications are a great example of this approach. Figure 7.11 presents a snapshot of a rich set of assessment specifications that integrate the content and skills that are valued in the classroom and discipline (National Assessment Governing Board, 2018). Note that each content area (columns) is crossed with the practice (rows), which gives rise to a differentiation in the

**Classification of Types of Assessment Targets in the Three Major Assessment Areas
According to the Practices for Technology and Engineering Literacy**

	Technology and Society	Design and Systems	Information and Communication Technology
Understanding Technological Principles	Analyze advantages and disadvantages of an existing Technology Explain costs and benefits Compare effects of two Technologies on individuals Propose solutions and alternatives Predict consequences of a Technology Select among alternatives	Describe features of a system or Process Identify examples of a system or Process Explain the properties of different Materials that determine which is suitable to use for a given Application or product Analyze a need Classify the elements of a system	Describe features and functions of ICT tools Explain how parts of a whole Interact Analyze and compare relevant Features Critique a process or outcome Evaluate examples of effective Resolution of opposing points of View Justify tool choice for a given Purpose
Developing Solutions and Achieving Goals	Select appropriate technology to Solve a societal problem Develop a plan to investigate an Issue Gather and Organize data and Information Analyze and Compare advantages and disadvantages of a proposed Solution Investigate environmental and Economic impacts of a proposed Solution Evaluate trade-offs and impacts of a proposed solution	Design and Build a product Using appropriate processes and Materials Develop forecasting techniques Construct and Test a model or Prototype Produce an alternative design or Product Evaluate trade-offs Determine how to meet a need by choosing resources required to Meet or satisfy that need Plan for durability Troubleshoot malfunctions	Select and Use appropriate tools to Achieve a goal Search media and digital resources Evaluate credibility and solutions Propose and Implement strategies Predict outcomes of a proposed Approach Plan research and presentations Organize data and information Transform from one Representational form to another Conduct experiments using digital Tools and simulations

FIGURE 7.11

Snapshot of Integrated National Assessment of Educational Progress Technology and Engineering Literacy Specifications

Note. ICT = Information and Communication Technology.

targets of measurement that exist at the intersection of content and skills, presenting the item writer with a more detailed set of expectations to guide their work.

Using RPLDs to Guide Assessment Design

The second area of integration that is a hallmark of PAD is using RPLDs as an essential element of assessment design. Recent reviews of testing programs' use of PAD (Huff et al., 2017) indicate that the use of RPLDs to guide design is still nascent at best, although using them has been long called for in designing assessments for which the primary purpose is to reliably classify students into a proficiency level and support valid inferences about what they know and can do (Bejar et al., 2007; Huff & Plake, 2010a; Luecht, 2013, 2019; Perie & Huff, 2016).

In conventional approaches to assessment design, articulation of PLDs of any sort does not occur during the design process; rather, the development of PLDs is considered the first step of the standard-setting methodology, which occurs after the

first test administration. In contrast, in PAD the items are designed to elicit evidence for a particular performance level (or, more specifically, for a particular claim about students where the claims are arranged along a performance continuum as expressed in the RPLDs).

This intentional, *a priori* match of an item to a specific performance level has several benefits. First, the claims about what students know and can do in each performance level are inferences that require an evidential argument that can be evaluated. To have reasonable confidence that inferences about what students know and can do are valid, this evidential argument must be compelling. Designing items from the outset that target specific performance levels seems to be an inherently plausible, commonsensical approach. The alternative argument—that items designed *absent* RPLD inputs will yield valid evidence helpful for classifying students based on what they know and can do—seems whimsical at best. A second, related benefit is that when items are designed to target specific RPLDs, we are essentially engineering the desired scale properties from the beginning, creating a coherence and integration between item design and psychometrics that is fragmented in conventional approaches to test design. Third, when items are targeted to specific claims about students organized as RPLDs, there is an opportunity to engineer cut scores rather than have standard setting be a completely *ad hoc* process (Huff & Plake, 2010b; Lewis & Cook, 2020).

Resistance to targeting items to specific performance levels is typically rooted in the observation that the intended progression of items does not turn out perfectly when reviewing the observed difficulty statistics for the item. For example, suppose there is an assessment that is intended to have three performance levels that are represented by the item response theory–based scale scores, as shown in the first column of Figure 7.12. Items were designed to align to performance levels, as shown in the second column. However, after data were collected from students in initial field trials and the items were analyzed and placed on the scale according to their estimated difficulty parameters (third column), there was a mismatch for many of the items. Some would use these mismatches as evidence that attempting to target items to specific performance levels is too imperfect of a science to be of utility. However, the mismatch of observed difficulty to intended difficulty is not a new problem—conventionally developed norm-referenced tests overdevelop item pools for precisely this reason, where crafting intended scale properties is mostly achieved through careful selection from an item pool after field testing, rather than designing the fit of items to scale intentionally from the beginning. But others argue—including the authors—that the process of targeting items to specific performance levels, reviewing data from field tests, hypothesizing on the cause of the mismatches, and revising items when needed—is simply best practice in the spirit of continuous improvement and iterative design (Kaliski et al., 2011; Lewis & Cook, 2020; Luecht, 2013, 2019; Schneider et al., 2013).

In the spirit of PAD, assessment designers have advocated seeing these mismatches as an opportunity for discussion and learning rather than dismissal of the entire process. Indeed, they contend that these discussions should occur in relation to all items,

not just the ones that are mismatched (Huff & Ferrara, 2010; Lewis & Cook, 2020). The following three questions should guide the evaluation of items: (a) Does the item placement along the scale support or detract from the assumptions as articulated in the domain model essential elements (approach to cognition and RPLDs) and the design patterns (task features, difficulty drivers)? (b) Do any assumptions or design features need to be revised? (c) What can we assume about opportunity to learn (OTL), and how does that influence the data for the items? For items that are intended to be rather simple in complexity and the statistics on the item support this intention, like Item 1 from Figure 7.12, it is usually assumed that most students have had OTL to learn the knowledge and skills tested by the item and that students in the *Basic* category have some barrier to OTL (e.g., do not have the prerequisite skills to access this grade-level content). These same types of discussions should happen with items at the other end of the spectrum, like Item 9 from Figure 7.12. The item was intended to be complex and the item location along the scale seems to support the assumptions that went into the RPLD and item design. However, the design team reviewing the item and resulting statistics should not discount the counterhypothesis that the item is complex as the result of some flaw (e.g., construct-irrelevant variance) or that our assumptions about the complexity of the target of measurement are flawed and the main reason that the item statistics are showing the item as hard is that very few students have had OTL. The role of OTL is especially critical to consider for items that are mismatched: All of the assumptions that go into the item design can be supported by research and expert judgment, the item can be free of flaws, and the item can still locate on the scale in an unexpected way because of a lack of OTL (making the item seem more complex than it is had students had the OTL for the targeted knowledge and skill), widespread

RPLD and Scale Score Range	Intended RPLD	Observed Placement on Scale
Basic 350–450	Item 1	Item 1
	Item 2	Item 4
	Item 3	Item 3
Proficient 451–550	Item 4	Item 2
	Item 5	Item 5
	Item 6	Item 8
Advanced 551–650	Item 7	Item 6
	Item 8	Item 7
	Item 9	Item 9

FIGURE 7.12

Hypothetical Comparison of Intended Performance Levels and Item Scale Placement

Note. RPLD = range performance-level descriptor.

exceptional instruction in a particular area (the targeted content and skill is very complex and sophisticated, but most students are getting the item correct because they have had effective instruction and OTL), and/or more nefarious reasons such as item exposure or cheating (making complex items seem easy).

The mismatch between items and intended performance level needs more attention because understanding these mismatches helps us make better assessments and, in turn, may help us better understand—and improve—the assumptions about the construct of interest, student cognition, the underlying performance continuum as expressed by the RPLDs. It is in this process that our assumptions are either confirmed or disconfirmed and thus revised. This is the messy business of reasoning from imperfect evidence that is at the heart of PAD.

Using PAD to Engineer Desired Scale Properties

Tests for which the primary purpose is to provide norm-referenced information, such as a percentile rank, or tests for which the primary validity evidence is the strength of the prediction from a score to a criterion, such as grade point average, do not require sophisticated psychometric specifications. Typically, for these kinds of tests, psychometricians give very little direction to item writers in advance other than general advice to make sure there is a range of difficulty in items. It is assumed that high-quality items will produce an appropriate range of item discrimination and very few items will be flagged for differential item functioning. The goal of the psychometrician for these kinds of tests is to keep reliability high and maintain the unidimensional scale over time as new items are added and the population shifts. Norms and predictions are monitored and updated on a routine but not annual basis. The scale is reset as infrequently as possible because maintaining historical trend is of utmost importance to the score user. In this context, the psychometric team will review test specifications upon development or revision in conjunction with the item-writing team to ensure that there are enough items of various characteristics (e.g., algebra or geometry, multiple-choice or short constructed response) to accomplish these goals. However, in conventional testing programs, psychometricians do not become heavily involved in the design of the assessment or items until there are field test data to analyze. As Luecht (2019, p. 8) lamented,

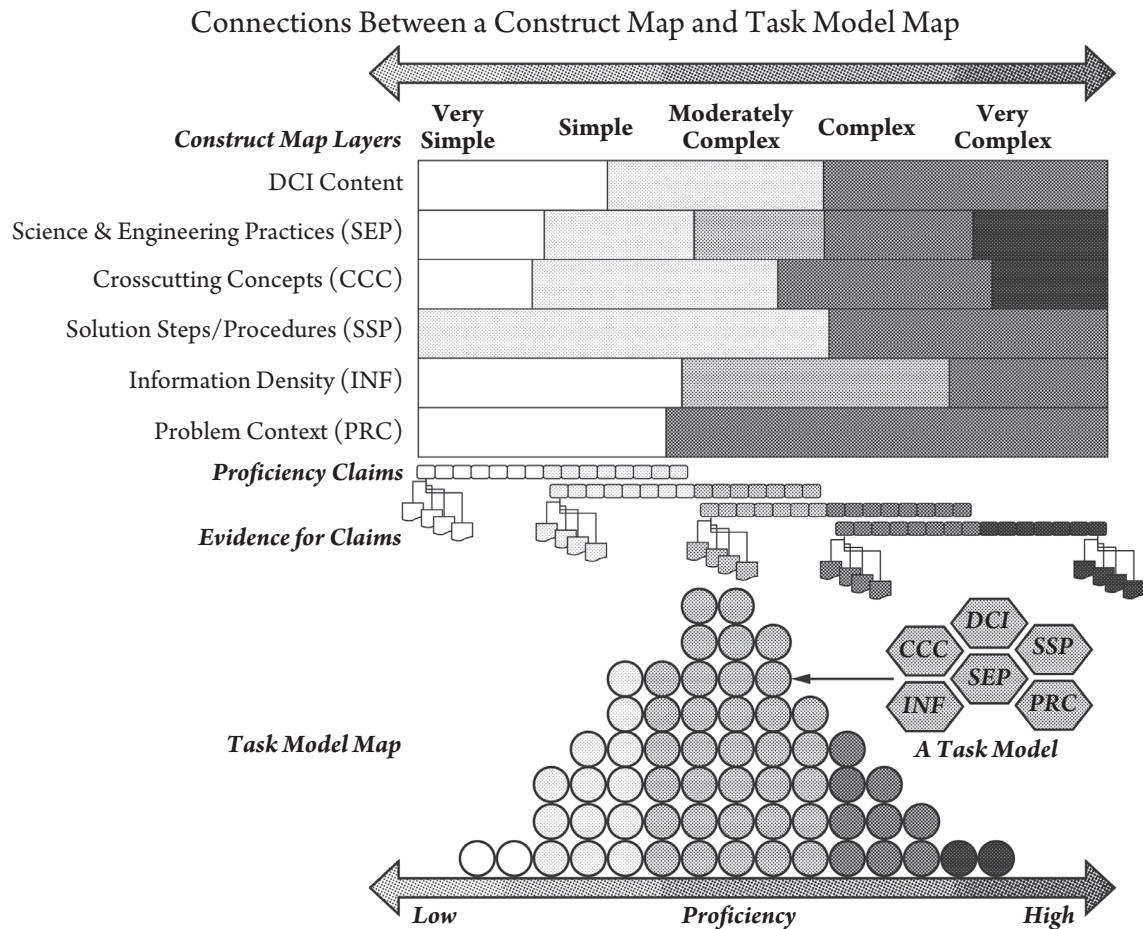
The fundamental dilemma seems to stem from the way that we design and implement operational testing programs and ultimately develop our score scales. Most test forms and the score scales are usually constructed in siloed procedures, carried out by different groups within an organization doing different things at different times without extensive coordination nor careful integration when it comes to validity and score use. [Item writers] tend not to worry about the statistical properties of the scale, psychometricians often view test content as an annoyance that gets in the way of optimizing certain statistical scale properties, and end users must then rely on PLDs and standard setting and ad hoc score reporting techniques to build some semblance of meaning into the score scales.

For tests that need to support inferences about student proficiency in regard to a set of learning standards, we argue that psychometricians and item writers must be close collaborators from the beginning and throughout the design process, as part of an interdisciplinary team with representatives from other, related fields, such as but not limited to classroom teachers and learning scientists, long before test specifications are developed or items are field tested. Through this collaboration, the intended scale properties can play a central role in the design process. Specifically, assessments that will require performance standards—or cut scores—so that each student can be classified into a proficiency level require sufficient measurement information to distinguish students who are, for example, *Basic* versus *Proficient* versus *Advanced*. For this reason, item writers need to change their perspective from writing items that span a wide range of difficulty to targeting the items to distinguish between performance levels, which requires the RPLDs to inform assessment design and item writing. The criticality of doing so must be understood by psychometricians and item writers alike for the scale to support a successful standard setting (Huff & Plake, 2010b; Lewis & Cook, 2020; Luecht, 2013, 2019) and for the resulting inferences to have compelling evidentiary argument that leads to validation of the inferences about what students in each performance level know and can do. The intentional relationships among our research-based assumptions about student cognition, the latent proficiency continuum as expressed in the RPLDs, task features, difficulty drivers, and the resultant scale properties of the assessment should result in a strong body of evidence to support valid inferences about what students know and can do.

One hypothetical example of what an integrated set of content and skills to be assessed and psychometric specifications could look like is provided by Luecht (2019) in Figure 7.13. In this example, Luecht used the NGSS to illustrate that when appropriately engineered, design patterns (see the term *task model* in Figure 7.13) will produce measurement information that is located on a deliberate range of the scale to support intended inferences. This can only happen when RPLDs—composed of claims about student proficiency that integrate skills and content and reflect a progression of student cognition—and the desired scale properties are understood by the full interdisciplinary team to be complementary components and are used as the basis of the assessment design from beginning to end.

SUMMARY

PAD helps us integrate and create coherence where conventional approaches to assessment design are fragmented. Three areas of integration were discussed in this section: first, test specifications that integrate content and skills in meaningful ways to reflect the hypothesized learning trajectory articulated in the RPLDs and serve as the basis for the eventual inferences about what students know and can do; second, the use of RPLDs as the basis for iterative item design and development; and third, the integration of item

**FIGURE 7.13**

Generating Performance-Level Descriptors for the Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores Partnership

Note. From “Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores (SCILLSS): The Role of Performance Level Descriptors for Establishing Meaningful and Useful Reporting Scales in a Principled Design Approach (White Paper)” by R. M. Luecht, January 2019, Nebraska Department of Education. https://www.scillsspartners.org/wp-content/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf

design and scale design. These three related levels of integration require collaboration across an interdisciplinary team from the beginning and throughout the iterative design and development process, especially by those responsible for designing and developing the items and those responsible for designing and maintaining the scale through psychometrics. It is this same partnership that is required to articulate a robust inferential argument that is supported by evidence and supports valid interpretations of what students know and can do.

The integrations discussed in this section have profound impact on how we think about the inferential and validation arguments. With PAD, we have the opportunity to develop a validation argument that is shaped by the deeper notion of coherence, rather than the typical superficial alignment, and includes evidence to support our intended

interpretations of student performance that is generated from the beginning of the design process, not post hoc validation studies.

THE EVIDENTIARY ARGUMENT IN PAD

One hallmark of PAD is that the evidentiary argument (which is defined as the evidence to support the inferential argument) and the validation argument (which is defined as an evaluation of how compelling the evidence supporting the inferential argument is) are central components of the iterative design endeavor, rather than a series of post hoc analytical exercises to be documented in a technical manual. For the purposes of this chapter, we posit that:

1. A Kanesian perspective on assessment validation (Kane, 2006, 2013) complements both a theory of action framework and the PAD framework.
2. A *theory of action* (TOA) is required to frame the evidentiary and validation arguments for assessment in the larger educational context.
3. The evidentiary argument and the validation argument are composed of a series of claims and evidence that are hierarchical, nested, and represented in many forms and in many grain sizes. For example, the *evidence model* in ECD is an evidentiary argument for specific targets of measurement, but only a subcomponent of the larger evidentiary argument supporting score interpretation, which in turn is only a subcomponent of the larger evidentiary argument supporting the effectiveness of score use as framed by the TOA.
4. The educational assessment industry needs to hold itself accountable for the decisions made throughout the assessment design process that either support or hinder the intended inferences about what students know and can do through the production of *procedural validity evidence for assessment design*.
5. Given this context, in PAD, many post hoc “validity studies” become obsolete. We use post hoc alignment studies as the prime example.

We discuss each of these propositions in turn in the following section.

Assessment Validation

As discussed previously, Toulmin’s (1958, 2006) thinking on the use of practical arguments, rather than formal logic, played an influential role in the evolution of PAD. The extension of the use of practical arguments to assessment validation processes has been an integral part of PAD from the outset. Kane (2013) described an argument-based approach to validation as requiring that

the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (p. 1)

In PAD, we think of the evidentiary argument as the evidence used to support the inferential chain of reasoning threaded throughout the design components of the assessment. What evidence do I have to support that the approach to cognition is appropriate for this domain? These grade levels? What evidence do I have that the RPLDs follow a reasonable progression of student learning from novice to mastery within a given grade and domain? What evidence do I have to support that the design patterns and resulting items will yield evidence of where along the latent proficiency continuum a student most likely resides? What evidence do I have that the resulting performance-level classifications support intended inferences given the stated purpose and use of the assessment? The answers to these questions—and many more like them—constitute the evidentiary argument. The validation argument is an evaluation of how compelling the evidentiary argument is. It is incumbent on the test publishers to provide the evidentiary argument. The potential test users determine the persuasiveness of the evidentiary argument by their decision to use or not use the test for its stated purpose and use. However, that decision does not constitute a “validation” of the test. Validation would require an external review of the evidentiary argument that evaluates the documentation of the design process and the articulated answers to all of the questions that arise through that process, as well as a determination of whether the evidentiary argument is, indeed, compelling and coherent.

Theory of Action

Since 1992, states have been mandated by the U.S. Department of Education to test, on an annual basis, certain subjects in certain grades for various accountability purposes and with some promise that these assessment results can also inform decisions at various levels of the educational system (federal, state, district, building, classroom). For example, when the No Child Left Behind Act became federal law in 2002 (No Child Left Behind Act, 2002) and extended the requirement from the previous law (Improving America’s Schools Act, 1994) that state assessment results be disaggregated by various subgroups, states used those data to make decisions about how to allocate funding for curricular supports, professional development, and other services. Districts also administer a variety of assessments throughout the school year for various purposes and uses, as do principals at the school level and classroom teachers. When thought through with purpose, these assessment endeavors complement each other in a balanced assessment system (Marion et al., 2019; Perie et al., 2009). When not coordinated as a system, this results in a proliferation of redundant testing that can have many negative, unintended consequences on teaching and learning. (See Ho & Polikoff, this volume, for additional discussion of test-based accountability in K–12 education.)

Bejar (2010), building on work from Bejar et al. (2007) and Mislevy and Haertel (2006), suggested that when PAD is employed, it is incumbent on the designers to consider both construct representation and convergent and discriminant validity evidence. Bejar (2010) highlighted the latter

because test scores, certainly those of high visibility, do not live in isolation and have consequences, intended and unintended, for students and their parents. Such consequences can be thought of as part of the evidence for and against a particular assessment. (p. 381)

Bennett (2015) further extended this notion of examining consequences by postulating that in a teaching and learning context, assessments are a form of treatment and the intended effects of the assessment must be examined through the lens of efficacy or TOA.

Figure 7.14 provides a sample TOA for an interim assessment for which the primary purpose and use are to help teachers identify gaps in student learning in a particular domain of interest (e.g., fourth-grade learning standards for measurement and geometry) and to support changes in instructional practice with the intended outcome of improving student learning. From this perspective, both the evidentiary argument and the validation argument are subsumed by a broader TOA for the role that the interim assessment plays in teaching and learning.

In this example, evidence to support fidelity of implementation is a key component of the TOA. For example, although the score reports may be optimally designed to inform instructional actions, without the appropriate support systems in place the reports may not be used as intended. One can imagine several systemic actions that would need to be evaluated: Was adequate professional development provided to

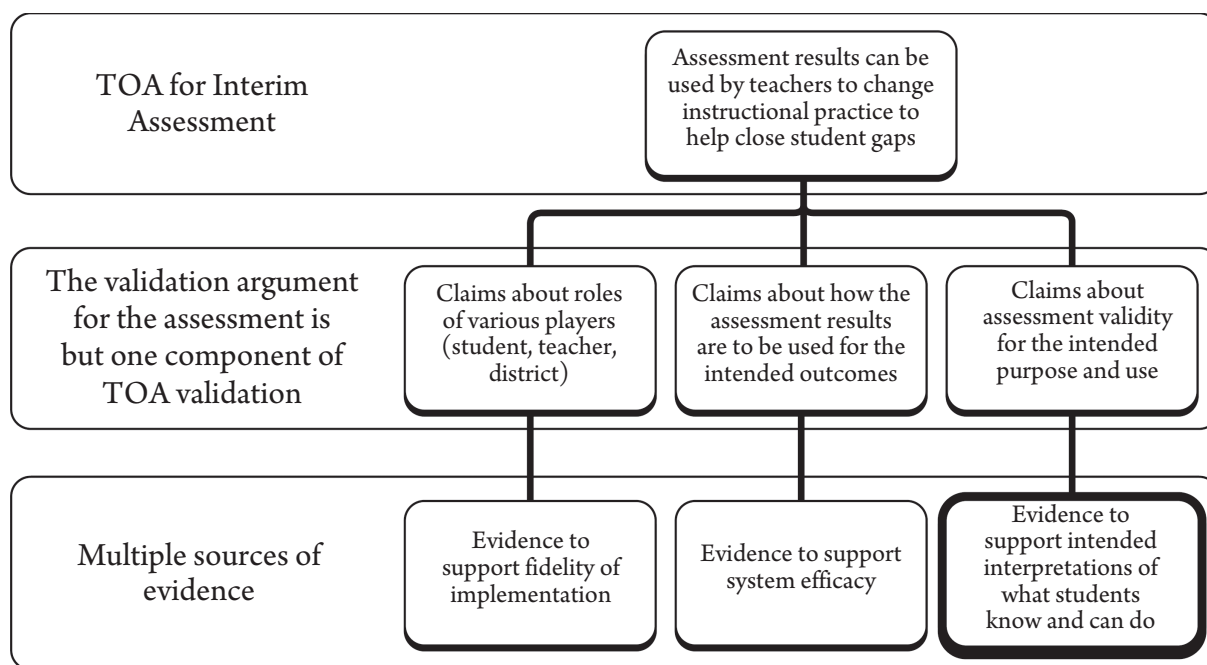


FIGURE 7.14

Sample Theory of Action for Interim Assessment

Note. TOA = Theory of Action.

support appropriate interpretation and use of the reports? Were teachers given timely access to the reports? Were teachers provided time to review the reports and plan accordingly? Do teachers have access to curricular and instructional supports that cohere with the assessment results? When not implemented with fidelity, it is unlikely that an assessment designed to adjust instructional practice will result in efficacious student learning gains.

Research that examines the impact of the interim assessment as a “treatment” should occur to determine whether the actions made based on the score report result in improved achievement for students (Brookhart, 2009; Nichols et al., 2009; Shepard, 2009). There is a growing body of research suggesting that teachers often struggle to determine next instructional actions to take (Heritage et al., 2009; Ruiz-Primo et al., 2010; Schneider & Andrade, 2013; Schneider & Gowan, 2013; Schneider & Meyer, 2012). In light of this finding, if a testing program’s claim that the intended purpose and use of the assessment is to inform instruction, this must be supported by assessment designers’ documented work with an interdisciplinary team—expert teachers, learning scientists, score report designers, to name a few—to develop instructional recommendations. To return to Figure 7.14, score interpretations must be supported by evidence that supports their accuracy (evidentiary argument) and instructional recommendations should be supported by evidence that demonstrates that faithful implementation of the recommendations results in student learning gains (efficacy research).

Most testing programs focus on the evidence to support score interpretation in the lower right box of Figure 7.14 to the detriment of other sources of evidence. Gathering and documenting evidence to support score interpretation is usually treated rather mechanically, as demonstrated by the routine structure of most technical manuals that include a predictable series of psychometric analyses. When PAD is used, we argue that a rich and coherent evidentiary argument is articulated by the very nature of the design process itself.

Claims and Evidence

Assessments are a type of evidentiary argument (Kane, 1992; Messick, 1989; Mislevy, 1994). As our claims about what students know and can do become more complex and the uses of assessment results become more varied and complex, the burden on our evidentiary argument increases. Simultaneously, as educational tests come under greater scrutiny, it is incumbent on test makers to ensure that the evidentiary argument supporting the inferences we are making about students is both clear and compelling. There are at least two ways that PAD differs from conventional test development that help tremendously with creating a clear, compelling evidentiary argument. First, the inferential argument is defined up-front, used as the basis of design, documented clearly throughout the design process, and updated during assessment development; this documentation serves as both design tools (i.e., the essential elements of PAD) and the evidentiary argument that supports the inferential argument. Second, one way that evidence is conceptualized in PAD is the

observable manifestation of what the student knows and can do, and the articulation of that which is observable is used as a design element throughout the assessment development process. Collecting and curating these elements of the evidentiary argument throughout the design process contrasts with conventional approaches to test development, in which “evidence” is narrowly defined as the collection of empirical data that can only happen when design is complete and scores are analyzed.

Under PAD, the term *evidence* is used in different contexts to reference various aspects of the larger body of evidence. Evidence is used broadly to refer to the evidentiary argument that captures the comprehensive set of claims that begins with the definition of the construct and ends with the use of the test score information to accomplish the intended purpose of the assessment. During design, development, and administration, these claims are supported by the collection and documentation of design process artifacts, including but not limited to the essential elements and empirical data. Psychometric evidence is collected to support claims that items are appropriately designed and scored to locate students along the performance continuum as articulated by the RPLDs. In addition, evidence also refers to the data, warrant, and backing supporting a claim of the TOA. Table 7.3 gives examples of how claims and evidence are used in nested levels within PAD as design features as well as components of the evidentiary argument.

In Figure 7.15, we show the hierarchical and nested nature of the claims that constitute the evidentiary argument for an examination program designed to award college credit and placement.

Procedural Validity Evidence

PAD, as described in the previous section, evolved from a rich history of the role of evidentiary reasoning in assessment; a natural consequence of engaging in PAD is the documentation and use of several layers of the evidentiary argument in the form of the various artifacts that represent the essential elements of PAD, for example, prioritized knowledge and skills in the domain (domain analysis), the approach to cognition and RPLDs (domain model), and the task features and difficulty drivers (design patterns). Nichols et al. (2017) posited that to achieve the intended purpose and use of an interim or summative assessment in the educational system, coherence must exist in the design process in at least two ways: (a) All design elements must be informed by theories of learning and cognition, and (b) procedural validity evidence must be collected to demonstrate that the assessment designers followed the process and used the design elements as intended. If the intended process was followed, then the procedural validity evidence will be natural artifacts of that process.

PAD starts with an assessment design plan that explicates in detail the processes that will be followed during design, the points of iteration, the artifacts that will be produced, and the key attributes of those artifacts. In this way, procedural validity evidence for the claim “The intended assessment design process was followed” can be collected and evaluated. Such a notion may seem novel in assessment design, but it is not. We

Table 7.3 Claims and Evidence in Principled Assessment Design

Grain Size	Claim	Example Evidence
Smallest: A single item with a particular <i>evidentiary focus</i> (i.e., target of measurement)	Student can distinguish among levels of abstraction in conflicting historical information	Task stimuli (historical text, map, and/or graphic) includes various levels of abstraction. Task prompt focuses student on distinguishing among levels of abstraction
Small: A single learning standard	Student can evaluate conflicting historical information for particular historical period and focus (e.g., pre-Columbian migration patterns)	Observable evidence has four components: • Recognition and response to conflicting information • Clear, comprehensive thesis • Significant depth and quantity of evidence to support thesis, including main concept and supporting details • Recognition of different levels of abstraction in historical information
Medium: A primary design component for the assessment (e.g., RPLDs or design patterns)	The RPLDs represent a research-based progression of proficiency that are useful both for assessment design and for making inferences about what students know and can do	Literature review on how students learn and build knowledge and skills in the domain of interest; data (survey, focus groups) that indicate educators are making appropriate inferences about what students know and can do from the PLDs; procedural validity evidence from the design pattern development and item-writing process that PLDs are being used as intended in design
Medium: One of the five sources of validity evidence cited in the <i>Standards</i> (AERA et al., 2014)	Students employ intended cognitive processes when responding to assessment items	Research report from a think-aloud study that indicates students are using the intended cognitive processes at each performance level
Large: Evidence that one or more claims in the TOA are supported	When teachers use the assessment results to drive instruction, students achieve higher gains in learning than when teachers do not	Quasi-experimental design study showing respectable effect sizes for claim that when teachers use the results of the assessment to drive instruction, students have higher gains than when assessment results are not used

Note. PLD = performance-level descriptor; RPLD = range performance-level descriptor; TOA = theory of action.

routinely articulate our intended methodologies and our rationales for every psychometric aspect of the test-making endeavor: establishing and maintaining the scale over time, equating or linking forms (or monitoring item drift), item analyses, standard setting, and postoperational validation studies (e.g., dimensionality analysis, predictive studies). These plans are meticulously detailed by the assessment provider, reviewed by independent technical advisors who provide feedback, and then reviewed and approved by the assessment client (e.g., the district or state). The assessment providers are held responsible for incorporating feedback from the advisors and client and then are held accountable for executing the psychometric plans and presenting detailed evidence of said execution. In addition, there are widely held expectations throughout the industry

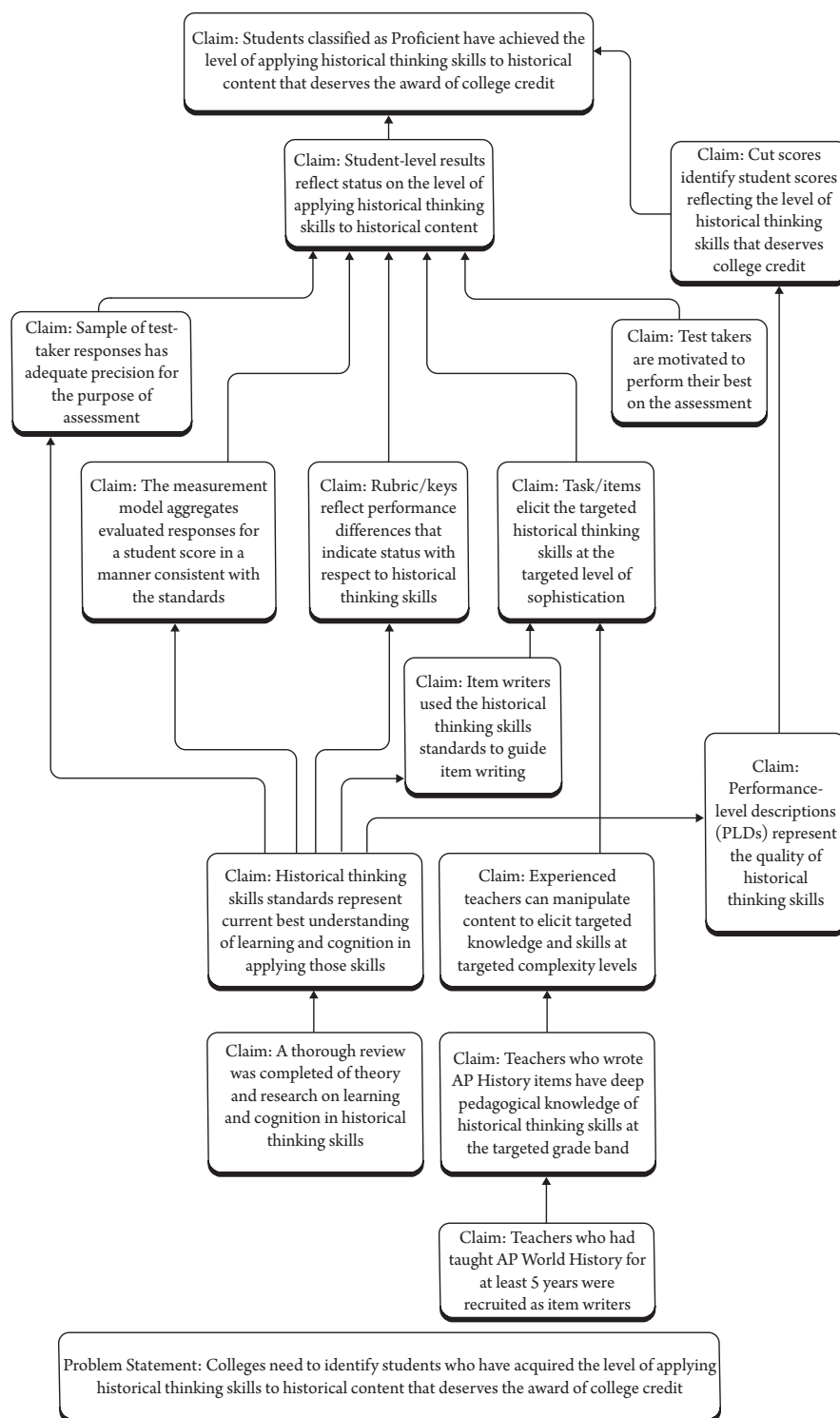


FIGURE 7.15
The Nested and Hierarchical Nature of Claims

on what constitutes acceptable and thorough documentation of these psychometric methods and analyses, which for federally mandated state summative tests are, in fact, externally audited via peer review. The whole of assessment design and development should be held to the same rigorous expectations of making a plan, executing it, and providing evidence that it was executed as intended.

An assessment design plan outlines the design and development process in detail, as well as what evidence will be produced that demonstrates that the plan was executed. Given that most of design and development relies on human judgment, assessment designers should be held accountable for providing procedural validity evidence. When PAD is engaged, procedural validity evidence is a natural by-product of the process as documentation of the assumptions and rationales undergirding each decision is used as design tools (e.g., the approach to cognition, design patterns). It is also helpful to be explicit in the plan on the role and nature of iteration, especially regarding the use of field test data to refine RPLDs and design patterns.

Table 7.4 contains a proposed set of validation criteria for PLD development expanded to include criteria for design patterns and item development. Huff and Plake (2010b) adapted these criteria from what are typically required of the standard-setting process. There is no reason that we should not apply the same high expectations to assessment design and development.

Hendrickson et al. (2013) offered two sets of checklists for use in the PAD endeavor. One is a set of criteria that can be used to evaluate design pattern quality. The criteria are based on assumptions that the items are written to discern between increasing levels of performance as articulated in the RPLDs and that there will be iteration of either or both RPLDs and design patterns after field testing. The second example from Hendrickson et al. (2013) is a checklist regarding iteration for PAD design components that relies heavily on iteration and consensus across various stakeholders on an interdisciplinary team. How do we know when we are done and can move forward to the next phase?

A final example in Figure 7.16 represents an item specification checklist to help ensure that the item writer is keeping critical design elements in mind. Notice how the checklist requires the item writer to indicate whether the item is to discern between students who are at the lower end of the proficiency continuum (Performance Levels 1 and 2: identify) or between students who are at the upper end of the proficiency continuum (Performance Levels 3 and 4: analysis) and how those items have different characteristics related to student cognition. It is easy to see from this example how a design tool can also serve as procedural validity evidence and evidence to support an inferential argument that makes explicit the relationships among approach to cognition, RPLDs, design patterns, items, and inferences about what students know and can do.

Alignment

Alignment of test specifications and item content to the learning standards is a key tenet of content validity for educational assessments. In conventional approaches to test design, alignment is conducted after item writing is complete. We contend that when

Table 7.4 Validation Framework and Criteria for Performance-Level Descriptors, Design Patterns, and Item Development

Category of Evidence	Criterion	Potential Evidence for PLDs
Procedural	Care in selecting panelists	Qualification and representation of panelists to support claims panel has expertise in how students learn and progress in domain and what it looks like as students reach proficiency and beyond
	Justification for PLD development framework	The approach chosen should be justified and the learning science paradigm made explicit
	Panelist training	Surveys indicate panelists understand the framework and have been sufficiently trained
	Appropriate contribution from panelists	Panelists' contribution, discussions, points of consensus, and compromise should be documented; panelist surveys
	Proper implementation	Documentation of implementation compared to workshop design: agenda, panelist surveys, PLD creation development templates; training framework
	Panelist confidence	Survey at end of panelists' confidence in process, quality, and successful implementation of development framework
Internal	Sufficient interpanelist consistency	Points of disagreement among panelists should be addressed and effectively moderated to achieve consensus or compromise
	Decreasing variability across rounds	Panelists' judgments should converge throughout the PLD development process via consensus building
	Consistency across independent panels	PLDs from two independent panels should have multiple points of consistency and few points of difference; document through qualitative comparison
	Consistency across panel subgroups	Reasonable panelist characteristics (e.g., teacher vs. researcher, assignment to different discussion groups during PLD session) should not impact
External	Expert reviews	Expert reviews should confirm evidence from learning sciences is found within PLDs; expert reviews of PLDs should confirm utility for item writing
	Teacher review	Resulting PLDs should be reviewed for interpretability by a separate panel of teachers or through a public review
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?
Procedural	Care in selecting designers	Qualification and representation of panelists to support claims that panel has expertise in how students learn and item design
	Justification for task model development framework	The approach chosen should be justified and the grain size of and intended design patterns for item types made explicit in cases where design patterns are not created for all item types
	Designer training	Surveys indicate that designers understand the intended goals and outcomes of design patterns and have been sufficiently trained

Category of Evidence	Criterion	Potential Evidence for PLDs
	Appropriate contribution from designers	Designer contribution, discussions, points of consensus, and compromise should be documented; designer surveys
	Proper implementation	Documentation of implementation compared to planned design: agenda, surveys, creation of design patterns; training framework
	Designer confidence	Survey to assess designers' confidence in process, quality, and successful implementation of development framework
Internal	Sufficient consistency	Points of disagreement among designers should be addressed and effectively moderated to achieve consensus or compromise
	Increased confidence across iterations	Iterative feedback should increase in perceived utility for consensus building
	Consistency across subgroups	Designer characteristics (e.g., teacher vs. researcher, assignment to different grade levels) should not impact quality of outcome
External	Expert reviews	Expert reviews should confirm that evidence from learning sciences is found within design patterns; expert reviews of design patterns should confirm utility for item writing
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?
Procedural	Care in selecting item writers	Qualification and representation of item writers to support claims item writers have expertise in the item writing and subject area and grade level assigned
	Item-writer training	Item writers are trained on the use of design patterns and show evidence of being sufficiently trained (e.g., can diagnose features not included in item that should be for a specific achievement-level target) before being allowed to develop items independently
	Proper implementation	Artifacts of training implementation including training framework, practice materials, feedback, and training recordings are collected
	Item-writer confidence	Survey at end of item-writer training to show confidence in process, quality, and readiness to move forward
Internal	Sufficient intraitem writer consistency	Item writer shows evidence of developing and/or identifying intended item features to match PLDs, including appropriate item types for descriptors, and receives feedback when needed
	Increased confidence across iterations	Item writers' feedback on perceived utility should increase as process is implemented
External	Expert reviews	Expert reviews should confirm evidence of training is sufficiently robust; expert reviews of prototype tasks should confirm utility for item-writer training
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?

Note. PLD = performance-level descriptor. Adapted from "Innovations in Setting Performance Standards for K–12 Test-Based Accountability," by K. Huff and B. Plake, 2010, *Measurement: Interdisciplinary Research and Perspectives*, 8(2), 130–144.

Grade 6- RL.3 <i>Describe how a particular story's or drama's plot unfolds in a series of episodes as well as how the characters respond or change as the plot moves toward a resolution</i> <i>Analyze how and why individuals, events, or ideas develop and interact over the course of a text.</i> Central aspect: • <i>Items measure students' ability to analyze how characters change as the plot moves toward resolution. Elements may also include how a plot unfolds in episodes</i>		
Item Measuring CCLS RL.3	Yes/No	If "No," Explain or Describe
Measures central aspect: (PL 1–2) The item requires identification of the change in a character as plot unfolds. Possible stems may include: Stem: How does character X change in lines XX–XX? Stem: Which of the following best describes the change in character X in lines XX–XX? Stem: What does line X reveal about a character? OR Measures supporting aspect: (PL 3–4) The item requires analysis of change or shift in plot Possible stems may include: Stem: The change/shift in lines XX-XX develops the plot by Stem: Which lines from the story show the character's change from X to Y?		
The item stem does not reveal: • the interaction of elements • the key change/development of characters Unless the interaction is identified in the stem intentionally		
The analysis in the item is supported by the text (i.e., there is development of story elements)		
The item requires students to comprehend the majority of the passage to answer the item correctly		
THE ITEM MEASURES THIS STANDARD		

FIGURE 7.16**Item Development Checklist Example**

Note. From “Large-Scale Standards-Based Assessments of Educational Achievement,” by K. Huff, Z. Warner, and J. Schweid, in A. A. Rupp & J. P. Leighton, (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 397–426), 2017, Wiley–Blackwell.

PAD is employed, post hoc alignment studies are largely obsolete because the evidence of alignment is produced as part of the design process.

For example, in PAD, items are designed to measure a specific performance level of a learning standard in the RPLDs, and this level is documented. It is best practice that after items are developed, they are reviewed by parties external to the design team, such as some combination of classroom teachers, content domain experts, pedagogy experts, accessibility experts, and experts in cultural and linguistic responsiveness. As part of their review process, the experts could indicate with which RPLD level the item best aligns. These data could be analyzed for interrater reliability. This alignment evidence can also serve as a way to identify areas where more work may be needed; for example, low interrater reliability estimates for particular items can spark the same types of discussions among the design team as when psychometric analyses reveal where along the scale the item locates and whether that location is as intended.

Post hoc alignment studies may still have a defensible role in instances where assessment items were originally designed to meet one set of learning standards, and because of cost and/or time constraints, the items need to be examined for alignment to a revised or different set of learning standards. This is especially true in the state testing context because learning standards tend to change frequently, in either superficial or meaningful ways, as a consequence of leadership changes and political reasons rather than in response to new learning science research. Most states require alignment studies that show that items designed for a previous or different set of learning standards can be used with confidence to assess their new or revised learning standards because assessment design is expensive, and in the K–12 context, the cost ultimately falls on the shoulders of taxpayers.

In addition, many states—even if alignment evidence via a PAD process is compelling—may still require an alignment study conducted by an independent evaluator using conventional alignment methodologies in the short term, such as Webb (2005). No matter whether alignment is engineered from the outset or a study is done post hoc, the notion of what constitutes alignment needs to broaden to one of coherence across all aspects of assessment design: learning standards, RPLDs, item design specifications, scale properties, and interpretations about what students know and can do.

Evidentiary Argument Summary

In summary, when we engage in PAD we are defining our inferential argument and collecting evidence for our evidentiary argument from the beginning of the design process, which makes creating and collecting evidence for our validation argument an ongoing endeavor, rather than a set of post hoc analyses. In PAD, evidence comes in a variety of forms and represents a broader conception of evidence than the conventional narrow definition that evidence is equivalent to the data that are collected from administered items and analyzed psychometrically. We assert that the assessment industry should hold the assessment design process to the same kinds of scrutiny that we do our psychometric analyses and other judgmental-based processes, such as standard setting.

ASSESSMENT DESIGN: LOOKING AHEAD

Assessment designers and measurement professionals in the early 21st century are extremely fortunate to have several challenging and interesting problems to solve with emergent innovations. As demand grows for more rapid feedback that supports valid inferences about students and is instructionally actionable, there is growing interest in “stealth assessment,” smart games, and other mechanisms by which to leverage technology to seamlessly integrate learning and assessment in engaging ways for students (DiCerbo, 2014; Hong et al., 2019; Kim & Shute, 2015; Petrusel, 2014; Shute & Ventura, 2013; Young et al., 2011).

On a related note, at the time of this writing, the proliferation of assessment products flooding the marketplace that use one or more families of models lumped into the broad category of artificial intelligence (AI) is staggering. The models undergirding these products are developed by and large without the benefit of any assessment design

or psychometric theory and practice. The ways in which this evolution will change the field and the industry are unknown. If that were not sufficiently challenging, there is a growing desire for novel constructs to be incorporated into these blended learning and assessment systems. The scientific practices of the NGSS (e.g., developing and using models, planning and carrying out investigations) and the ACT Holistic Framework for Education and Work Success (e.g., collaborative problem-solving, sustaining effort) are examples of increasing complexity of learning and assessment needs. As the complexity of targets of measurement increases, the case for using PAD becomes stronger. It is hard to imagine a compelling validity argument without a transparent chain of inference that links all of the essential components together—for example, targets of measurement, RPLDs, items, scale properties, intended score inferences—coherently and elegantly.

Three related challenges will need to be addressed before we can realize the full potential of integrated learning and assessment and the potential of AI to revolutionize learning and assessment writ large. As we will argue, these issues cannot be addressed, much less solved, without a commitment to PAD. First is the use of accessibility features and the relationship of these features to the targets of measurement. Second is the evolution from ensuring that assessments are fair for all students and devoid of bias to assessments that reflect and value the diversity of cultures that our students represent. The third issue is the emergent focus on student motivation and engagement as it relates to estimates of their proficiency. The role of student motivation and engagement in assessment is made especially complex when considered alongside accessibility features for students with a variety of disabilities, as well as culturally and linguistically responsive assessments for students from our historically marginalized populations. These complex relationships need to be examined particularly closely for assessments that may not carry the same gravitas for students as assessments that contribute to their grades, promotion, placement, or other stakes that are important to students, such as interim assessments, assessments embedded in instruction, or other game-based or stealth assessments.

Accessibility and Cultural and Linguistic Responsiveness

It is beyond the scope of this chapter to give full treatment to the topic of accessibility, accommodations, and fairness or, as fairness is treated with a more contemporary perspective, cultural and linguistic responsiveness, in educational assessment. Other chapters in this volume are dedicated to a full treatment of these topics (see Zwick and Rodriguez & Thurlow). These issues are addressed here in light of the challenges that remain to be addressed in assessment design and development to ensure our assessments are accessible and responsive to the cultural and linguistic diversity of the students whom we serve. That said, the research and practice in these areas are developing rapidly and whatever is articulated here in 2025 is likely to be out of date in the next year or so. Nonetheless, we include examples here of how, without the precision and transparency demanded by the PAD process, incorporating accessibility and culturally and linguistically responsive features could jeopardize the validity of the inferences about what students know and

can do. As we have demonstrated in this chapter, PAD forces the detailed articulation of the assumptions about the intended targets of measurement and the intended inferences, and the resulting design specifications can be used to support decisions about accessibility and culturally and linguistically responsive features such that inferences about students from different populations are as free of interference as possible.

Let us begin with an example of an item from a standards-based, adaptive, computer-based interim assessment whose primary purpose and use is to inform instruction through categorizing students at the most beneficial point along an instructional pathway given their strengths and weaknesses in the target domain. The example item in Figure 7.17 illustrates an item designed with keen attention to both the National Center on Educational Outcomes recommendations to ensure that universal design for learning is applied to assessments and the WCAG (W3C, 2018) guidelines for accessibility features for items delivered via the Internet. These guidelines include but are not limited to simple, clear, and intuitive instructions and procedures, maximized readability and comprehensibility, and maximized legibility (National Center on Educational Outcomes, n.d.). Other features from WCAG guidelines include the following:

- strong color contrast between the shape and its background
- additional black outlining around the edges of the shape for added visual definition
- bold font used for answer selection; color contrast between black font and white background
- the mathematical expression is not an image so that an automated screen reader can read it aloud
- alternative text is included for this image so that a screen reader can be used for students who are visually challenged; for example, “A rectangular prism that is shaped like a shoebox. A legend indicates that $l = 7$, $w = 3$, and $h = 2$.”
- keyboard navigation allows the student to use the keyboard arrow keys to move from element to element in the correct order; therefore, a student using a screen reader will hear the many elements in this order and can navigate forward and backward between elements as needed:
 - the stem
 - the expression
 - the alt text describing the prism and the variable values
 - the four answer choices in the order shown on screen
- Spanish transadaptation

This example of accessibility features is rather straightforward. However, the line between accessibility and infringement on the intended target of measurement can blur, for example, with the addition of an audio option for reading aloud the text, item prompt, and response options that is included for many assessment items and available to all students—even those who do not have difficulty processing written text or visual impairment. An audio option likely would not interfere with the intended target

The expression shows the surface area of a rectangular prism with length l , width w , and height h . What is the value of the expression when the variables have the given values?

$$2lw + 2wh + 2lh$$

$l = 7$
 $w = 3$
 $h = 2$



78 82 72 84

AL_L6_677
Answer: Correct

Done →

La expresión muestra el área total del prisma rectangular con la longitud l , el ancho a y la altura h . ¿Cuál es el valor de la expresión cuando las variables tienen los valores indicados?

$$2la + 2ah + 2lh$$

$l = 7$
 $a = 3$
 $h = 2$



82 84 78 72

AL_L6_677
Answer: Correct

Acabé →

FIGURE 7.17

Example Item in English and Spanish to Illustrate Accessibility Features

of measurement for a straightforward math item like the one shown above, but would present a conflict for items that are designed to measure, for example, reading comprehension (as opposed to comprehension regardless of receptive mode). Without rigorous interrogation and documentation of the assumptions undergirding each assessment design choice, such as the kind demanded by PAD, we risk sloppy measurement and, as a result, flimsy arguments supporting our claims about what students know and can do when accessibility features are part of the item design.

Items are designed to be fair to all students (e.g., to avoid terms that privilege particular students when the concept or term is not the target of measurement) and are typically reviewed by independent panels to help ensure the items are free of sensitive topics or content (e.g., item-writing guidelines typically indicate that certain topics that could upset students be avoided, such as hurricanes or death) and bias (e.g., items do not include images or terms that portray students from marginalized communities in stereotypical or derogatory ways). At the time of this writing, our conventional notions of what constitutes fairness, sensitivity, and bias are being deeply interrogated as the assessment industry catches up with what pedagogy, curriculum, and instructional scholars and practitioners have known for quite some time: being responsive and inclusive of students' cultural and linguistic diversity matters in the learning endeavor (Gay, 2000; Hammond, 2015; Paris & Alim, 2017). In short, to be authentically culturally and linguistically responsive, we must rethink what is construct relevant and irrelevant and what is and is not sensitive to various populations (Randall, 2023; Solano-Flores, 2023). Hollie (2018) noted that Hollie (2012) defined culturally and linguistically responsive teaching as "the validation and affirmation of the home (indigenous) culture and home language for the purposes of building and bridging the student to success" (p. 23). In this context, culture has a broad definition and meaning; Hollie's work defines culture along multiple dimensions: ethnicity, sexual orientation, nationality, socioeconomic status, religion, gender, and age. For example, gardens are a favorite context for math items. Rather than having the context be a backyard (typical of suburbs) for every item that uses a garden, make sure that there is a broader spectrum of representation that includes farms (rural areas), community gardens (urban areas), and gardens from other cultures (e.g., terraced gardens from Indonesia). Another more poignant example would be to ensure that our assessment passages do not erase the lived experiences and histories of students from marginalized communities. For example, in a reading comprehension assessment where the passage topic is, for example, the 1893 World's Fair, a passage that only celebrates the wonder and achievements of the fair but fails to mention that luminaries Frederick Douglass and Ida B. Wells protested the exclusion of African Americans from exhibits (Duster, 1970) erases the history, and therefore the culture, of Black students in classrooms today. As we expand our contexts to be more responsive to and reflective of our ever-diversifying student body, we will need to

grapple with the inherent tension of attending to culturally and linguistically responsive content and contexts while simultaneously adhering to accessibility and sensitivity guidelines that steer item writers away from anything that may be potentially context rich or controversial. These debates must occur with a shared understanding of the target of measurement and what constitutes construct-relevant and construct-irrelevant variance. As these discussions become more nuanced, the precision, clarity, and transparency that are the hallmarks of PAD become more and more needed.

Engagement and Motivation

The research base on the role of student engagement and motivation in learning and assessment continues to grow. Unless students are engaged in the assessment, they will not be motivated to perform their best, and the assessment results for those students are likely underestimates of what they know and can do. This is a primary concern for many who are questioning what the lack of culturally and linguistically responsive assessments has meant for the achievement results of our students from marginalized populations (Gutiérrez, 2017; Lyiscott, 2019; Randall, 2021, 2023; Solano-Flores, 2019, 2023). For interim assessments designed primarily to inform instruction, rather than to assign grades, lack of engagement and motivation can be a real issue (Wise & DeMars, 2005). As most K–12 testing is slowly but surely moving toward computer-based rather than paper-based formats, and as the promise of blending learning and assessment to become the same endeavor unfolds, assessment design is in need of partnership with user experience designers (UX designers, UX design). User experience, in this usage, encompasses the various elements of design expertise that go into creating effective items, including interaction design, user interface design, art creation and curation, and usability.

Experts in UX design bring a perspective to assessment design that is generally not represented in interdisciplinary teams of psychometricians, content experts, educators, item writers, and learning scientists. Conventional practice would rely on illustrators or photo editors to participate in a very narrow way, such as providing a graphic as part of an item stimulus after an art specification has been defined by the item writer, which is a questionable practice given that said art specifications could benefit greatly from trained UX designers. However, UX designers, when part of the team from the start of the assessment design process, can help us think through the student experience of the assessment in new and compelling ways. The basic goals of UX design are to create an experience that identifies and meets the needs of the user. Some of the hallmarks of a well-designed experience are simplicity, transparency of goals and actions, clear communication of information, and an experience that is enjoyable to use. These goals translate to the experience of student assessment as well. There are more points of entry for UX design to partner in a computer-based test than a paper-based test, but UX designers also have perspective and expertise to offer in paper-based testing, since the goals and principles of good communication design apply to all media. In computer-based testing, UX designers are concerned

with not only the graphics required for items, but also the student navigation experience—from starting the assessment, to moving between items, to the within-item experience. The goal is to ensure that everything presented to the student and each part of the navigation process is intuitive and does not get in the way of the intended purpose: to optimally measure what the student knows and can do. In other words, the UX designer is just as committed to avoiding construct-irrelevant variance as the item writers and psychometricians. UX designers help assessment designers avoid construct-irrelevant variance through probing the assumptions that undergird item design with questions like:

- Are the interactions clear and easy to use? Are they age appropriate?
- Is the art content accessible and equitable? Does it represent the word, object, or concept in an unambiguous way the student can understand?
- Are there superfluous, decorative, or distracting elements within an artwork that may mislead or inhibit a student?
- Does the experience feel familiar and consistent from item to item?
- Are items designed in such a way that stimulus, stem, distractors, and answer areas are consistent across item types so that a student needs little time to figure out how to input their answer?
- Does the UX support taking the assessment over one session or multiple sessions?
- Does the user have appropriate context for what they are doing? Do they know why they are doing it?
- Does the hierarchy of content and images within an item inform the student as to where and how to answer the question?

With a PAD approach to assessment design, these types of questions would be considered in the development of the design patterns and the assessment delivery model. A strong partnership with the UX design team will help make these models better.

CHAPTER SUMMARY

Industry Inflection Point for PAD

What might be the future for the use of PAD? Andy Grove, Intel's former chief executive officer, observed, "When spring comes, snow melts first at the periphery, because that is where it is most exposed" (McGrath, 2019, p. 14). Might the same be true of the assessment industry? We might find the future influences on PAD not in the large-scale admissions tests or the state summative assessments, but in the game-based and stealth assessments, the assessment of emerging constructs like three-dimensional science learning, and the assessment of noncognitive constructs such as social-emotional learning.

PAD has been available to the field of educational assessment for over 20 years (Mislevy et al., 2002). Prior to the work of Mislevy and colleagues, the foundational concepts

for PAD were described by other scholars working at the intersection of learning, cognition, and assessment, including but not limited to Lohman and Ippel (1993), Nichols (1994), Embretson (1998), Pellegrino et al. (1999), and Snow and Lohman (1989). The use of PAD appears to offer many benefits to assessment developers and users. Yet, PAD does not enjoy widespread use in operational educational assessment design and development. Pieces have been implemented here and there in organizations and testing programs. For example, a version of PAD (ECD) was used in the design of the National Board for Professional Teaching Standards (Pearlman, 2008a) and the Test of English as a Foreign Language (Pearlman, 2008b). It was also used in the redesign of the AP program (Huff & Plake, 2010a) and by the Cisco learning network (Behrens et al., 2010), and it was required by the 2009 Race to the Top legislation that supported the assessment consortia (U.S. Department of Education, 2010). PAD is also being used by organizations to design assessments to measure the NGSS (Harris et al., 2019; Luecht, 2019). However, at the time of this writing, the authors are unaware of any operational large-scale assessment programs that systemically implement all the essential elements of PAD. The lack of widespread PAD adoption suggests that funders and users currently perceive less value in PAD compared to conventional test development.

The popularity of PAD may be waiting for an assessment industry strategic inflection point to dramatically increase PAD's value to assessment developers and users. A strategic inflection point is a change in the business environment that throws some assumptions into question and upends the basic assumptions of a business model. When this happens, we posit that the perceived value of PAD will outweigh the perceived value of conventional assessment development. Conventional assessment development and psychometric practices yield a sufficient return on investment for test developers. What is needed is a large-scale counterexample that demonstrates that PAD maximizes return on investment in assessment development.

Leading indicators—things that are not yet undisputed facts in an industry—suggest that the assessment industry is creeping toward an inflection point for PAD. A leading indicator for this inflection point is that the constructs of interest are becoming more complex and less approachable using conventional assessment development (Nichols & Huff, 2017). In addition, educators and policy makers in states and districts are increasingly unhappy with both the amount of testing and the quality of educational assessments. As such, we expect educational assessments to simultaneously meet multiple purposes and uses and to be more transparent with regard to their quality and meaningfulness in teaching and learning. Finally, assessment designers are under pressure, and rightfully so, to make sure our assessments are accessible to all students, are culturally and linguistically responsive, and are engaging to boot. The testing industry has the power of technology on its side in meeting these challenges, especially as it becomes easier for interdisciplinary teams to collaborate remotely. In addition, learning science research is occurring at a rapid rate that helps us design assessments that better support inferences about what students know and can do. Taken together, maybe there is an inflection point on the near horizon.

The arrival of inflection points is difficult to predict. For example, Reed Hastings, the founder of Netflix, waited and waited for the inflection point for the online streaming model. As Hastings explained,

In 1997, we said that 50% of the business would be from streaming by 2002. It was zero. In 2002, we said that 50% of the business would be from streaming by 2007. It was zero. . . . Now streaming has exploded. . . . We were waiting for all these years. Then we were in the right place at the right time. (McGrath, 2019, p. 79)

Similarly, Mislevy et al. (2002) were anticipating an inflection point in the testing industry over 20 years ago:

Standard procedures for designing and carrying out assessments have worked satisfactorily for the assessments we have all become familiar with over the past half century. Their limits are sorely tested today. The field faces demand for more complex inferences about students, concerning finer grained and interrelated aspects of knowledge and conditions under which this knowledge can be to bear. Advances in technology can provide far richer samples of performances, in increasingly realistic and interactive settings; how can we make sense of this complex data? And even with familiar assessments, cost pressures from continuous testing and social pressures for validity arguments demand more principled assessment designs and operations. (p. 126)

The philosophical orientation, practices, and procedures of PAD must be fully embraced and operationalized by each organization and each individual within who is responsible for designing and developing assessments of educational achievement, whether blended with instruction, interim, or summative. Our students, parents, teachers, principals, district staff, state educational leaders, and the public whom we serve deserve our best efforts to support the educational endeavor. Conventional approaches to educational assessment do not represent our best thinking and our best work; we can do better and we must.

ACKNOWLEDGMENTS

The authors would like to thank the editors and reviewers for their thoughtful comments and questions that made this chapter—and our thinking—crisper and clearer. Special gratitude goes to our mentors and pioneers in assessment design Mari Pearlman, Linda Steinberg, Ric Luecht, Steve Ferrara, and Bob Mislevy. We are grateful to Eli Mintzer for formatting the figures.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist*, 22(4), 297–306.

- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world* (CRESST Report 778). University of California, National Center for Research on Evaluation, Standards, and Student Testing. <https://files.eric.ed.gov/fulltext/ED520431.pdf>
- Bejar, I. I. (2010). Application of evidence-centered assessment design to the Advanced Placement redesign: A graphic restatement. *Applied Measurement in Education*, 23(4), 378–391. <https://doi.org/10.1080/08957347.2010.510969>
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1–30). JAM Press.
- Bennett, R. E. (2015). The changing nature of educational assessment review of research in education. *Review of Research in Education*, 39(1), 370–407.
- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.
- Brookhart, S. M. (2009). *Exploring formative assessment*. Association for Supervision and Curriculum Development.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.
- CAST. (2018). *Universal Design for Learning Guidelines*, version 2.2. <http://udlguidelines.cast.org>
- College Board. (2019). *AP [course]: Course and exam description*. <https://aphighered.collegeboard.org/courses-exams/course-exam-redesign>
- Council of Chief State School Officers & Association of Test Publishers. (2013). Operational best practices for statewide large-scale assessment programs. <https://ccsso.org/sites/default/files/2023-09/ATP%20Best%20Practices%20Version%202-FINAL-082113-interior-300dpi.pdf>
- DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society*, 17(1), 17–28.
- Duster, A. M. (Ed.). (1970). *Crusade for justice: The autobiography of Ida B. Wells*. The University of Chicago Press.
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 300–396.
- Ercikan, K., & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking*. Routledge.

- Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education*, 23(4), 325–341.
- Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). National Council on Measurement in Education and American Council on Education.
- Ferrara, S., Dogan, E., Glazer, N., Gorin, J., Haberstroh, J., Hain, B., Huff, K., Larkin, J., Nichols, P., Piper, C., & Sheehan, P. (2014, April 3–7). Development of cognitive complexity measures for PARCC (Presentation). In A. Rupp (Chair), *Cognition and assessment SIG business meeting and poster session*. American Educational Research Association Annual Meeting, Philadelphia, PA, United States.
- Ferrara, S., & Steedle, J. (2015, April 15–19). *Predicting item parameters using regression trees: Analyzing existing item data to understand and improve item writing* [Paper presentation]. National Council of Measurement in Education Annual Meeting, Chicago, IL, United States.
- Ferrara, S., Steedle, J., & Kinsman, A. (2015). *PARCC cognitive complexity: Analysis 1, 2, and 3 results*. Partnership for Assessment of Readiness in College and Careers. <https://eric.ed.gov/?id=ED599050>
- Frederiksen, J. R., & Collins, A. (1998). A systems approach to educational testing. *Educational Researcher*, 18(9), 27–32.
- Gay, G. (2000). *Culturally responsive teaching: Theory, research, and practice*. Teachers College Press.
- Gillmor, S. C., Poggio, J., & Embretson, S. (2015). Effects of reducing the cognitive load of mathematics test items on student performance. *Numeracy*, 8(1). <https://doi.org/10.5038/1936-4660.8.1.4>
- Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36(9), 923–936.
- Goldstein, D. (2020, February 15). An old and contested solution to boost reading scores: Phonics. *New York Times*. <https://www.nytimes.com/2020/02/15/us/reading-phonics.html>
- Gutiérrez, R. (2017). Political conocimiento for teaching mathematics: Why teachers need it and how to develop it. In S. Kastberg, A. M. Tyminski, A. Lischka, & W. Sanchez (Eds.), *Building support for scholarly practices in mathematics methods* (pp. 11–38). Information Age Publishing.
- Hammond, Z. L. (2015). *Culturally responsive teaching and the brain*. Corwin Press.
- Hanford, E. (2018, October 26). Why are we still teaching reading the wrong way? *New York Times*. <https://www.nytimes.com/2018/10/26/opinion/sunday/phonics-teaching-reading-wrong-way.html>
- Hanford, E. (2019, December 5). There is a right way to teach reading, and Mississippi knows it. *New York Times*. <https://www.nytimes.com/2019/12/05/opinion/mississippi-schools-naep.html>

- Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, 38(2), 53–67.
- Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K. W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications*. SRI International. https://cadrek12.org/sites/default/files/Harris_Krajcik_Pellegrino_McElhaney_constructing%20assessment%20tasks%202016.pdf
- Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis*. Council of the Great City Schools. <http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf>
- Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology*, 14(14).
- Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education*, 23, 358–377.
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative classroom assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31.
- Herman, J. L. (2010). *Coherence: Assessment success key to next generation* [AACC report]. University of California. <https://files.eric.ed.gov/fulltext/ED524221.pdf>
- Hollie, S. (2012). *Culturally and linguistically responsive teaching and learning: Classroom practices for student success* (1st ed.). Shell Education.
- Hollie, S. (2018). *Culturally and linguistically responsive teaching and learning: Classroom practices for student success* (2nd ed.). Shell Education.
- Hong, J. C., Chang, C. H., Tsai, C. R., & Tai, K. H. (2019). How situational interest affects individual interest in a STEAM competition. *International Journal of Science Education*, 41(12), 1667–1681.
- Hong, J., & Lissitz, R. W. (Eds.). (2017). *Test fairness in the new generation of large-scale assessment*. Information Age Publishing.
- Huff, K., Alves, C. B., Pellegrino, J., & Kaliski, P. (2013). Using evidence-centered design task models in automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 102–113). Routledge.
- Huff, K., & Ferrara, S. (2010, June). *Frameworks for considering item response demands and item difficulty* [Paper presentation]. Council of Chief State School Officers National Conference on Large-Scale Assessment, Detroit, MI, United States.
- Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press.

- Huff, K., & Plake, B. (2010a). Evidence-centered assessment design in practice. *Applied Measurement in Education*, 23, 307–309.
- Huff, K., & Plake, B. (2010b). Innovations in setting performance standards for K–12 test-based accountability. *Measurement: Interdisciplinary Research and Perspectives*, 8(2), 130–144.
- Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, 23, 310–324.
- Huff, K., Warner, Z., & Schweid, J. (2017). Large-scale standards-based assessments of educational achievement. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 397–426). Wiley–Blackwell.
- Hurford, D. P., Hurford, J. D., Head, K. L., Keiper, M. M., Nitcher, S. P., & Renner, L. P. (2016). The dyslexia dilemma: A history of ignorance, complacency and resistance in colleges of education. *Journal of Childhood & Developmental Disorders*, 2(3). doi: 10.4172/2472-1786.100034
- Improving America's Schools Act of 1994. Pub. L. No. 103–382 (1994).
- Johannesson, P., & Perjons, E. (2014). *An introduction to design science*. Springer International Publishing.
- Johnstone, C., Altman, J., Thurlow, M., & Moore, M. (2006). *Universal design online manual*. University of Minnesota, National Center on Educational Outcomes. <https://nceo.info/Resources/publications/UDmanual/UDmanualPrint.htm>
- Kaliski, P., Huff, K., & Barry, C. (2011, April). *Aligning items and achievement levels: A study comparing expert judgments* [Paper presentation]. Meeting of the National Council on Measurement in Education, New Orleans, LA, United States.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger Publishers.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education*, 87, 340–356.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798.
- Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction*. Lawrence Erlbaum Associates.
- Lane, S., Raymond, M., & Haladyna, T. (Eds.). (2016). *Handbook of test development* (2nd ed., pp. 119–143). Routledge.

- Lawrence, I. M., & Shea, E. C. (2008). *Improving assessment: The intersection of psychology and psychometrics* (ETS Research Memorandum No. RM-08-15). ETS.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26, 3–16.
- Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.
- Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39(1), 8–21.
- Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Lohman, D. F., & Ippel, M. J. (1993). *Cognitive diagnosis: From statistically based assessment toward theory-based assessment*. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 41–71). Lawrence Erlbaum Associates.
- Louisiana Department of Education. (2020). *A teacher's guide to LEAP 360*. <https://www.louisianabelieves.com/docs/default-source/assessment/a-teachers-guide-to-leap-360.pdf?sfvrsn=4>
- Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, 14(1), 1–38.
- Luecht, R. M. (2019, January). *Strengthening Claims-Based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS): The role of performance level descriptors for establishing meaningful and useful reporting scales in a principled design approach* [White paper]. Nebraska Department of Education. https://www.scillsspartners.org/wp-content/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf
- Lyiscott, J. (2019). *Black appetite. White food: Issues of race, voice, and justice within and beyond the classroom*. Routledge.
- Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019). *The challenges and opportunities of balanced systems of assessment: A policy brief*. National Center for the Improvement of Educational Assessment. <https://files.eric.ed.gov/fulltext/ED598421.pdf>
- McGrath, R. (2019). *Seeing around corners: How to spot inflection points in business before they happen*. Houghton Mifflin Harcourt.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–23.

- Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). American Council on Education.
- Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439–483.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). American Council on Education.
- Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). SRI International.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Lawrence Erlbaum Associates.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–67.
- National Assessment Governing Board. (2018). *Technology & engineering literacy framework for the 2018 National Assessment of Educational Progress*. <https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/technology/2018-technology-framework.pdf>
- National Assessment Governing Board. (2021a). *Mathematics framework for the 2026 National Assessment of Educational Progress*. <https://www.nagb.gov/naep-subject-areas/mathematics/2026-naep-mathematics-framework.html>
- National Assessment Governing Board. (2021b). *Reading framework for the 2026 National Assessment of Educational Progress*. <https://www.nagb.gov/naep-subject-areas/reading/framework-archive/2026-reading-framework.html>
- National Center on Educational Outcomes. (n.d.). *Universal design of assessments: Overview*. Retrieved September 14, 2020, from https://nceo.info/Assessments/universal_design/overview
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. <http://www.corestandards.org/read-the-standards/>
- National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. The National Academies Press. <https://doi.org/10.17226/6160>
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. <https://doi.org/10.17226/10019>
- National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. The National Academies Press. <https://doi.org/10.17226/10129>

- National Research Council. (2005). *How students learn: History, mathematics, and science in the classroom*. The National Academies Press. <https://doi.org/10.17226/10126>
- NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academic Press. <https://www.nextgenscience.org/search-standards>
- Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, 64(4), 575–603.
- Nichols, P., Ferrara, S., & Lai, E. (2016). Principled design for efficacy: Design and development for the next generation of assessments. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common Core Standards, Smarter-Balanced, PARCC, and the nationwide testing movement* (pp. 49–81). Information Age Publishing.
- Nichols, P., & Huff, K. (2017). Assessments of complex thinking. In K. Ercikan & J. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments: The use of response processes* (pp. 63–74). Routledge.
- Nichols, P., Kobrin, J. L., Lai, E., & Koepfler, J. D. (2017). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications, first edition* (pp. 13–40). Wiley–Blackwell.
- Nichols, P., Meyers, J., & Burling, K. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice*, 28(3), 14–23.
- No Child Left Behind Act of 2002. Pub. L. No. 107–110, § 115, Stat.1425 (2002).
- Paris, D., & Alim, H. S. (2017). *Culturally sustaining pedagogies: Teaching and learning for justice in a changing world*. Teacher’s College Press.
- Pearlman, M. (2008a). Chapter 3: The design architecture of NBPTS certification assessments. In R. E. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards: Advances in program evaluation* (Vol. 11, pp. 55–91). Emerald Group Publishing.
- Pearlman, M. (2008b). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). Routledge.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and instructional practice. *Research in Education*, 24, 307–352.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2011). *Knowing what students know: The science and design of educational assessment*. National Academies Press. <https://www.nap.edu/catalog/10019/known-what-students-know-the-science-and-design-of-educational>
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments, *Educational Psychologist*, 51(1), 59–81.

- Penuel, W. R., & Shepard, L. A. (2017). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). Wiley–Blackwell.
- Perie, M. (2006). *Convening an articulation panel after a standard setting meeting: A how-to guide*. Center for Assessment. https://www.nciea.org/publications/RecommendforArticulation_MAP06.pdf
- Perie, M., & Huff, K. (2016). Determining the content and cognitive demand for achievement tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 119–143). Routledge.
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5–13. <https://doi.org/10.1111/j.1745-3992.2009.00149.x>
- Petrusel, R. (2014). Integrating click-through and eye-tracking logs for decision-making process mining. *Informatica Economica*, 18(1), 56.
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.
- Randall, J. (2023). It ain’t near ‘bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment*, 28(2), 68–82.
- Ruiz-Primo, M. A., Furtak, E. M., Ayala, C. C., Yin, Y., & Shavelson, R. J. (2010). Formative classroom assessment, motivation, and science learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative classroom assessment* (pp. 139–158). Routledge.
- Rupp, A. A., & Leighton, J. P. (Eds.). (2017). *The handbook of cognition and assessment: Frameworks, methodologies, and applications*. Wiley–Blackwell.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). National Council on Measurement in Education and American Council on Education.
- Schneider, M. C. (2017, April 26–30). *Using principled assessment design to support formative assessment and students’ opportunities to learn* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Antonio, TX, United States.
- Schneider, M. C., & Andrade, H. (2013). Teachers’ and administrators’ use of evidence of student learning to take action. *Applied Measurement in Education*, 26(3), 159–162.
- Schneider, M. C., & Gowan, P. (2013). Investigating teachers’ skills in interpreting evidence of student learning. *Applied Measurement in Education*, 26(3), 191–204.
- Schneider, M. C., Huff, K., Egan, K., Gaines, M., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement level descriptors. *Educational Assessment*, 18(2), 99–121.
- Schneider, M. C., & Johnson, R. L. (2018). *Using formative assessment to support student learning objectives*. Routledge.

- Schneider, M. C., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and mathematics. *Journal of Multidisciplinary Evaluation*, 8(17), 1–24.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.
- Schum, D. A. (2009). Science of evidence: Contributions from law and probability. *Law, Probability, and Risk*, 8, 197–231.
- Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, 28(3), 32–37.
- Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press.
- Simon, H. A. (1969). *The sciences of the artificial*. Massachusetts Institute of Technology.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). American Council on Education/Macmillan.
- Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education*, 4, 43. <https://doi.org/10.3389/feduc.2019.00043>
- Solano-Flores, G. (2023). How serious are we about fairness in testing and how far are we willing to go? A response to Randall and Bennett with reflections about the *Standards for Educational and Psychological Testing*. *Educational Assessment*, 28(2), 105–117.
- South Carolina Department of Education. (n.d.). *Adoption list of formative assessments*. <https://ed.sc.gov/tests/middle/adoption-list-of-formative-assessments/>
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.
- Toulmin, S. E. (2006). Reasoning in theory and practice. In D. L. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin model. New essays in argument analysis and evaluation* (pp. 25–29). Springer.
- U.S. Department of Education. (2010). *Race to the Top program guidance and frequently asked questions*. Authorized under Sections 14005 and 14006 of the American Recovery and Reinvestment Act of 2009. <https://www2.ed.gov/programs/racetothetop/faq.pdf>
- Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal*, 115(2), 270–289.
- van Aken, J. E., & Romme, G. (2009). Reinventing the future: Adding design science to the repertoire of organization and management studies. *Organization Management Journal*, 6(1), 5–12.
- Webb, N. L. (2005). *Web Alignment Tool (WAT): Training manual* (Draft Version 1.1). Wisconsin Center for Education Research, Council of Chief State School Officers.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

- Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, 33(3), 1–12.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1–17.
- World Wide Web Consortium. (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. <https://www.w3.org/TR/WCAG21/>
- Young, V.M., House, A., Wang, H., Singleton, C., & Klopfenstein, K. (2011, May). *Inclusive STEM schools: Early promise in Texas and unanswered questions* [Paper presentation]. Highly successful STEM schools or programs for K–12 STEM education: A workshop, Washington, DC, United States. https://sites.nationalacademies.org/cs/groups/dbasseite/documents/webpage/dbasse_072639.pdf

NOTES

1. We believe that classroom assessments could also benefit from principled assessment design principles and practices, but that is beyond the scope of this chapter.
2. For the purposes of this chapter, we are addressing students who take generalized assessments. Alternate assessments for students with the most significant cognitive disabilities are discussed elsewhere in this volume, including by Lane and Marion (in relation to validity) and Rodriguez and Thurlow (in relation to fairness).
3. For the purposes of this chapter, *task* and *item* are used interchangeably.
4. This chapter uses the term *item writer* to refer to all contributors to item development; however, in practice item writer generally refers specifically to those who write initial drafts of items. Most of the work attributed to item writers in this chapter is performed by content development professionals who shepherd those initial item drafts through the full item development process, which includes an iterative cycle of review and editing. These educational measurement professionals often have titles like *content specialist* or *editor*.
5. “Off the shelf” is an industry term that refers to tests developed for use in any state, regardless of the learning standards in the state.

Test Administration and Scoring

Mark D. Shermis

Performance Assessment Analytics

Allan S. Cohen

University of Georgia

Jordan M. Wheeler

University of Nebraska–Lincoln

The first four sources of validity evidence described in the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association [AERA] et al., 2014) focus on test content, response processes, internal structure, and relations to other variables. Each of these sources is heavily dependent on test administration and scoring. During test administration, the test taker interacts with the test; the presentation of the items should reflect the test content as intended by the test developer. And the test-taker interaction with the test should support the expected response processes. Item scoring should also align with the expected scoring behavior by the test developer and enable the analysis of data in support of validity. Ultimately, test administration and scoring procedures should support the intended interpretation of test scores. When test administration or scoring problems occur, they can undermine the validity of score inferences.

This chapter focuses on current and emerging practices in test administration and item-level scoring. We start by addressing the role of test administration in the assessment process and how to maintain the integrity of test administration through test security. Then, we discuss the role of human and machine-based scoring systems, provide examples of several machine-based scoring systems, and illustrate new technologies for determining text and speech content. Finally, we provide research results on machine scoring systems, suggest how to frame issues of reliability and validity for machine scoring systems, and reference how this technology is making inroads internationally.

TEST ADMINISTRATION

Four aspects of test administration support the validity of test score inferences: (a) the role of standardization, (b) test administration modalities, (c) test accommodations, and (d) nonstandard test administration.

The Role of Standardization

Standardization means that the test administration and scoring are applied consistently across all test takers. As noted in the *Standards*, “When directions, testing conditions, and scoring follow the same detailed procedures for all test takers, the test is said to be standardized” (AERA et al., 2014, p. 111). Standardization is intended to provide all test takers the same opportunity to demonstrate their competence and ensure that the interpretation of score(s) is the same for all test takers taking the same or different forms of the test.

The *Standards* (AERA et al., 2014) also take into account, however, that not all test takers may be able to demonstrate their competencies using the same procedures. Therefore, proper interpretation of standardization acknowledges that the same testing conditions may not be appropriate for all test takers. For example, test takers who receive extended time or have someone read the items to them as a test accommodation do not have the same testing conditions as test takers who do not receive these accommodations. In such cases, it is critical to specify, support, and ensure testing procedures

that provide all test takers with an opportunity to demonstrate the same knowledge, skills, abilities, or other competencies concerning the construct being measured. That is, ensuring the same construct is measured under such necessary differences in testing procedures is a significant concern. It is the responsibility of the test developer and administrator to anticipate such differences and to analyze their impact on the interpretation of test scores (see, in this volume, Huff et al., Zwick, and Rodriguez & Thurlow).

Standardization is intended to reduce construct-irrelevant variance due to differences in test administration procedures. For example, suppose familiarity with equipment or items used in the test administration is not part of the measured construct but may impact test takers' ability to demonstrate their competency. The principle of standardization requires that test takers be familiarized with the equipment or items prior to taking the examination. Further, it is necessary that those administering the test be familiar with the equipment or items and be trained to provide test takers with experiences that familiarize them with the equipment or items. Even when testing programs have attempted to make performance comparable across different testing environments, such as paper-and-pencil versus computer administration, meaningful differences may persist (Mazzeo & Harvey, 1988).

If appropriate variations in testing procedures cannot be supported, then it is possible the test may not measure the same construct(s) across test takers, thereby affecting the interpretation of results. The test developer and administrator need to anticipate such differences and, where necessary, provide appropriate interpretation modifications. Under the conditions for which this cannot be done, the test developer needs to note that standardization is likely to be affected; differences in interpretation should be described in the construct(s) being measured and interpretative materials should be provided recognizing such differences.

Test Administration Modalities

The typical steps involved in administering a group-based standardized assessment include some or all of the following: assigning test takers to tests; securing testing space, materials, and staff; staff training including proctoring and test administration procedures; checking in and seating of test takers; storing electronic or paper exam materials before administration; administering the test; monitoring and proctoring of test takers, including managing breaks and emergencies; checking test takers out from their session; irregularity reporting; distributing, collecting, and storing electronic or paper exam materials following the examination, and test item scoring and test scoring (Cohen & Wollack, 2006). Until recently, most large group testing programs used some type of paper-and-pencil administration. Computer-based testing, while not a new field (e.g., Jensema, 1974; Lord, 1977; Schmidt et al., 1978) has more recently begun to be implemented in an increasing number of large-group standardized testing programs. Developments in the statistical foundations for computer-adaptive testing (e.g., S. Wang et al., 2019), for example, have resulted in increased use in large-scale testing (S. Wang et al., 2016), as has the availability of computers and Internet access.

Despite the differences in test administration modality, however, there are variations in administering a large-group computer-based testing program compared to those in previous paper-and-pencil testing programs. These variations may come in the form of longer testing windows, standardization across multiple testing platforms, training of test administrators, use of technology-enhanced and -enabled test items, automated tools around accommodations, dynamic forms design via adaptive testing or linear on the fly, more sophisticated cheating detection algorithms, the development of procedures to handle situations where testing is interrupted (i.e., electrical outages), computer-based scoring and distributed scoring, and faster reporting. Most large-scale testing programs have a computerized version of their assessment (Mills et al., 2005). Importantly, the use of computer-based testing supports innovative item types and thereby the ability to assess test takers against the construct in ways that paper-based methods cannot support; these item types may better assess the intended constructs and expand the constructs being assessed. The use of innovative item types may align with the increased use of computers in test taker's daily lives as well as increase test takers' engagement and interest in the test. This may in turn result in test results that better reflect test takers' knowledge, skills, and abilities.

Monitoring of test takers has typically been done in person. However, the testing industry is increasingly relying on technology to assist with proctoring because it provides the ability to expand the pool of proctors and because it offers additional tools to detect potential cheating behavior. Some testing organizations (e.g., Prometric Testing, Pearson VUE) use video cameras for monitoring test takers. Video cameras allow in-person proctors and off-site proctors to monitor test takers, thereby expanding proctors' access to the test administration process. Technology is increasingly being used to monitor and record test takers' responses and to identify potential test security concerns, such as impersonation, accessing unauthorized materials, item theft attempts, or communicating with others (Dendir & Maxwell, 2020; Weiner & Hurtz, 2017).

Testing Accommodations

Issues regarding testing individuals who cannot be fairly tested under the usual administration conditions are covered by Zwick and by Rodriguez and Thurlow (this volume). In this chapter, we briefly discuss the necessity for recognizing that one set of administration conditions may not necessarily provide a fair opportunity for all individuals to demonstrate their competence. For accommodations to successfully level the playing field, in other words, they must be administered according to best practice as defined by research. This chapter focuses on the most common testing accommodations and refers to Zwick and Rodriguez and Thurlow for a fuller discussion.

Testing accommodations are provided to test takers as assistance in overcoming the effects of disabilities that are not part of the measured construct (Sireci et al., 2003, 2005; Thurlow et al., 2006). Corrective eyeglasses, for example, are an accommodation, although they are rarely considered a testing accommodation. Yet, when a test taker with poor vision is tested without corrective glasses, the measured ability is likely to be

confounded with construct-irrelevant variability due to poor vision. When a necessary accommodation is not provided, construct-irrelevant variability will confound the interpretation of scores to the extent of the impact of the disability. It is important to note that the set of accommodations typically depends on the subject and construct being measured, and the accommodations assigned to an individual test taker can differ based on the subject.

Extended testing time is the most commonly provided test accommodations (Gregg & Nelson, 2010; Lovett, 2010) and one of the most widely studied (Zenisky & Sireci, 2007). Other accommodations such as Braille, large print, sign language interpretation, and having questions read to a test taker are also commonly provided (Bolt & Thurlow, 2004).

Guidelines for the amount of extra time given to accommodate test takers have been provided in the Americans With Disabilities Act (U.S. Government Accountability Office, 2011). How the provision of extra time is handled during test administration is also covered by these guidelines. For example, extended time accommodation for tests using conventional paper-and-pencil format may require separating those receiving the accommodation into a different testing space so that proctors can easily keep track and monitor the students who received extra time. In contrast, computer-based tests can sometimes provide accommodation without moving the test taker to a separate testing room since computer-based tests can incorporate the appropriate amount of time for each student and exit the test once the allowed time is reached.

English language learners with limited English proficiency typically receive some type of accommodation when the language of the test is different from their native language (J. W. Young & King, 2008). J. W. Young and King noted, however, that there does not appear to be a commonly accepted set of accommodations (see also Rodriguez & Thurlow, this volume). One adaptation that is sometimes made in this regard is to provide a version of the test in the native language of the test taker (Buxton et al., 2014; Lazarin, 2006). This may be accompanied by either a translated version of the test or a dictionary that translates words in the test into the native language. This kind of accommodation assumes that the test taker will be able to understand the question if it is stated in their native language. One concern with this kind of accommodation, however, is that it could introduce construct-irrelevant variance because the language used for instruction is typically different from the language used for the test, which could potentially cause confusion if ideas and concepts cannot be directly translated.

Differential performance of test takers with and without accommodations may be caused not necessarily by the accommodation itself, but by the ability or level of competency of the test takers (Cohen et al., 2005). Evaluating the assessment of the construct needs to be done under each of the different administration conditions for which interpretations of the score are to be made. That is, it needs to be shown that the same construct is measured under each of the administration conditions. The purpose of the evaluation for each of the different administration conditions is to ensure

the comparability of scores obtained from the two administration conditions. For example, the scores produced by students who are given extended time need to be evaluated and compared to the scores produced by students who are not given extended time. Evaluating the effect accommodations have on measuring the construct is a crucial component of establishing test validity and fairness. A variety of methods are used to assess the effect of accommodations, including score-based prediction, differential prediction, differential validity, and criterion-related validity (see Cleary, 1968; Marini et al., 2019; Saka et al., 2022). Ultimately, test developers and test users are responsible for determining the impact of nonstandard procedures on the measurement of the underlying construct.

TEST SECURITY

Test security and detection of cheating are essential in ensuring fairness and supporting the validity of test scores. Violations of test security typically fall into four areas: (a) cheating by test takers and administrators, (b) exposure/disclosure of secure test material, (c) inappropriate test preparation, and (d) inappropriate test administration (Fremer & Ferrara, 2013). Of course, prevention is the first step in maintaining good test security. Having written policies, training test administrators, positively identifying test takers, and creating a secure test environment go a long way toward limiting cheating. In the next section, we briefly review current efforts at detecting threats to test security and different types of cheating.

Statistical Approaches to Detection of Cheating

Generally, a well-prepared test taker should not be surprised at the content on a test (Stroud, 1946). However, cheating could easily arise if the test taker feels unprepared for the test. Cheating is intended to provide the test taker with an unfair advantage, resulting in an inaccurate assessment of the test taker's proficiency (Cizek & Wollack, 2017). Cheating is not new and has been a concern in testing programs for over 2,000 years (DuBois, 1970). In ancient China, which is most probably where large-scale examinations were first administered, test takers used to write answers to questions on undergarments and refer to them during the examination. Most of the research on cheating has focused on high-stakes tests, with somewhat less attention paid to cheating on lower stakes tests.

Wollack and Schoenig (2018) described statistical detection of three types of cheating: (a) answer copying and collusion, (b) preknowledge, and (c) test tampering. Important in all methods for detecting cheating is the consideration of error rates and the recognition that statistical evidence alone is insufficient to support convicting a test taker of cheating. Despite a substantial and ongoing effort at studying methods for detection of cheating, there is no consensus on how to investigate potential misconduct on tests, particularly as it relates to the inclusion of statistical evidence. In this regard, it is crucial to understand the necessity for reporting error rates and aggregating

across multiple sources of evidence rather than sanctioning test takers solely based on statistical information.

Wollack and Schoenig (2018) described five broad categories of cheating detection: (a) person fit indices, (b) copying detection and similarity, (c) score differencing, (d) erasures and answer changes, and (e) response time methods. Cheating detection using person fit indices involves a statistical approach to identifying test takers who do not fit the psychometric model well (e.g., incorrectly answering items the psychometric model expects them to answer correctly and correctly answering items the psychometric model predicts them to answer incorrectly). Copying and similarity detection involve flagging test takers who have suspiciously similar response patterns. Score differencing utilizes compromised test items by identifying test takers whose performance differs between the subset of compromised and uncompromised items. Erasures and answer changes involve identifying test takers who change their answers more often than expected. Finally, response time methods use models to determine the necessary amount of time to answer an item and identify test takers with irregular response times.

Research on person fit indices has tended to show poor power in areas such as detecting answer copying. As a result, relatively little research on cheating has focused on these indices. An alternative to person fit is to compare the similarity of response strings for possible answer copying or collusion. Some of these methods compare only incorrect responses, and others compare the complete response string. In addition to detecting cheating among test takers, some work has been reported using response similarity and integrity of testing rooms or facilities. One caveat with similarity analysis is that it can be challenging to distinguish colluding test takers and test takers who have received the same instruction. Erasure detection methods are only applicable to paper-and-pencil tests, although answer changes have been studied for paper-and-pencil and computer-based tests. Computer-based testing also offers the possibility of providing log-file information that can accurately indicate answer changes.

Response time, initially studied for use in improving item parameter estimates (van der Linden, 2007), has also been studied for possible use in detecting cheating, as in the form of item preknowledge (Meijer & Sotaridona, 2006). One approach using this technique considers the residual estimated response time compared to the actual response time. The assumption in the use of response time is, in part, that a speedy response, given the item and ability characteristics, may indicate item preknowledge.

One variant of preknowledge is inappropriate access to test content. This occurs when test takers harvest or access harvested items from a previous administration of the examination. It can happen when test takers simply write down questions they remember after taking a test and then either use them to prepare for a retest or provide them to other test takers who are preparing for the same or an alternate form of the test. This latter action is considered theft. This is also the case when it occurs as plagiarism (see section “Machine Learning Approaches to Detection of Cheating”). One manifestation of harvesting items occurs when test takers share testing materials in some electronic format, such as from chat rooms or group social media accounts. In

addition to statistical methods for detecting cheating, other approaches such as web crawls, hotlines, auditing test centers, and logging and monitoring systems also have been used (Cizek & Wollack, 2017). These kinds of methods are sometimes helpful to detect some types of cheating, such as harvesting.

Machine Learning Approaches to Detection of Cheating

Some research is beginning to report the use of machine learning methods to detect cheating. Response similarity is one area in which this approach is being reported. Much of the early work on the use of machine learning for cheating detection began in the context of online games (Galli et al., 2011). Cheating in online games is a concern because games can often provide a substantial income stream for game developers (Yeung & Lui, 2008). Blackburn et al. (2014) surveyed STEAM, a large digital game delivery platform, and reported that approximately 6% of the 12 million users had been flagged as possible cheaters. STEAM developed and supports a cheating detection component called valve anti-cheating (“Valve Anti-Cheating Software,” 2020). Cheating is a concern in online games because honest players leave the game site when cheating is present (Alayed et al., 2013).

The increasing use of technology in education and society, in general, has led to the development of serious games (Abt, 1970). Game-based learning provides an important new platform for instruction and testing (Walker & Engelhard, 2014). Serious educational games developed out of this effort (Annetta, 2008) and contain testing components used to assess the game performance of the learner (Annetta & Bronak, 2011). Although some approaches have begun to be developed for detecting cheating in serious educational games (Rosyid, 2018), this is still a relatively new research area that is needed to address the security issues involved.

Recent work on cheating has focused on detecting plagiarism in constructed response items, most notably writing samples. Plagiarism detection is of interest because it is a form of answer copying. It differs from academic research in that use is made of another person’s work (e.g., writing, research) without appropriate attribution, whether done with or without modifying the actual words but retaining the meaning (zu Eissen & Stein, 2006). The possibilities for plagiarism have increased substantially over the years because of the very large number of machine-readable documents available on the Internet that can be used as sources (Alzahrani et al., 2012).

The availability of machine scoring (see section “Machine-Based Scoring Systems”) makes inclusion of constructed response items increasingly accessible for large-group testing. In this section, we consider recent work in machine learning that has focused on detecting plagiarism in documents that are in a machine-readable form. A number of these methods are available via online web applications, some of which are free and some which are not. Because the list of available methods for plagiarism detection is continually changing as machine learning and natural language processing methods evolve, we do not present a list of available methods. However, we suggest that a simple search on the Internet will yield multiple sites with plagiarism detectors.

Plagiarism detection methods generally use some form of document comparison against a database of documents, often using machine learning. In an early work on machine-based plagiarism detection, Shivakumar and Garcia-Molina (1995) described a word frequency method, which compared a new, possibly plagiarized document against a known record. These methods tend to work well when the comparison is made of directly copied text (zu Eissen & Stein, 2006). The use of intelligent plagiarism, whereby portions of the original text have been modified to retain the same meaning but with different words, is more difficult to detect (Alzahrani et al., 2012). Mozgovoy et al. (2010) described an application of a topic model to identify documents that contain a large amount of paraphrasing.

A careful consideration of error rates can often be missing from plagiarism detection sites. For example, academic work in a discipline may use the exact words and phrases as previously published research. Journals sometimes use plagiarism detection software to ensure that manuscripts accepted for publication do not plagiarize the work of others. Common wording is characteristic of much research that focuses on the same problem area; however, plagiarism detection methods that make use of common words or strings of words to detect plagiarism can sometimes erroneously result in a finding of plagiarism. As we suggested, software using older plagiarism detection methods can sometimes be circumvented by changing the wording or ordering of concepts.

In the next section, we focus on scoring systems, beginning with human scoring and then moving on to machine-based scoring systems.

HUMAN-BASED SCORING SYSTEMS

Complex items, such as constructed-response and performance-based tasks, often require human raters to score each response following a rubric or guideline. These scores are either assembled with other test scores, provided to the test taker, or used to train automated scoring engines so that not all responses need to be seen and scored by humans. The scores given to these items, however, are subjective by nature since raters need to judge the quality of the responses. When human raters are used to score these types of items, the raters become an essential component for ensuring the reliability, validity, and fairness of scores (McClellan, 2010). Human raters, therefore, often go through rigorous training protocols and methods to ensure that scores are accurate and reflect the construct being measured by the item (Glazer & Wolfe, 2020). Rater training protocols help assessment systems reduce construct-irrelevant variance among scores and increase the consistency of scores across raters (see ETS, 2021; Ricker-Pedley et al., 2020).

The level of rigor applied to rater training methods and protocols depends on the implications of the assessment results. High-stakes assessments often require a rigorous training protocol and different operational raters to score the same response to reduce construct-irrelevant variance, maintain high interrater reliability, and ensure the validity of scores.

General Steps for Rater Training Methods and Protocols

Rater training methods and protocols differ among systems, assessments, and items. These differences are based on multiple factors and include things such as the item type(s), the number of score categories (dichotomous or polytomous), the number of test takers, the context (i.e., low stakes versus high stakes), and the number of resources available. Developing a successful rater training protocol requires careful consideration of the intended use of the assessment because the training protocol will become part of the validity argument for the complex items.

The rigor of rater training methods and protocols depends on the implications of the assessment results. High-stakes assessments often require detailed training protocols and different operational raters to score the same response to reduce construct-irrelevant variance, maintain high interrater reliability, and ensure the validity of scores within and across test administrations (if applicable). However, the rater training methods and protocols follow a few general characteristics/steps (Wolfe & McVay, 2010). These methods and protocols are generally led by experienced raters and involve iterations of facilitated reviews about the correct application of scores. First, raters are provided information regarding the item, prompt, and related material(s). Next, raters review the scoring rubrics and anchor responses that have been annotated and scored. The anchor responses are typically created by content/domain experts for all possible score categories or selected from responses during the pilot phase of an item. Content and domain experts spend considerable time reviewing and discussing anchor responses since these responses serve as validity evidence and reference materials for scoring future responses. The qualifying exams are assessed by comparing the degree of raters' exact agreement and adjacent agreement with expert raters (e.g., 70% exact agreement and 100% adjacent agreement). Typically, raters must complete the training method and protocol to become an operational rater for a particular assessment or item (Bernardin & Buckley, 1981).

Materials and Information Regarding Prompt

Background information about the prompt and item is provided to the raters to ensure that they have a sufficient amount of information regarding the construct(s) being measured by the item. The information provided should demonstrate the breadth and depth of the construct and give raters an idea of the scope of the construct being measured. In addition to information and materials about the construct, raters are provided with materials that may be present in the item, such as source passages, figures, and tables. It is expected that raters fully understand the prompt and any supplemental materials, what is being asked, what construct is being measured, and how provided materials relate (and should be used) to the construct and scores.

Overview of Scoring Rubrics and Examples

Raters are provided with the scoring rubrics for the items and details about each of the score categories. Furthermore, raters are provided with example responses, often called anchor responses, to illustrate each score category on the rubric. These examples

provide adequate detail about the different aspects of the response and how they affect scoring. It is essential that the example responses not only provide information regarding each score category but also provide a wide variety of responses that illustrate the range of quality within a scoring category. Additionally, example responses that are confusing or unscorable are provided to prepare raters for identifying aberrant and unusual answers. These example responses serve as references, or anchors, for the raters as they score new responses.

The raters view the example responses, annotations, and score categories and discuss them, typically in a group setting, with a trainer or an expert rater. This process can take a considerable amount of time and requires strong facilitation skills by the trainer. It is at this stage where raters are being *calibrated* against the rubric.

Practice Scoring Responses

Once raters thoroughly understand the materials related to the prompt, the rubric, and example responses, they apply their trained scoring knowledge to practice responses. The set of practice responses covers all score categories and represents the range of responses within each score category. Each practice response has been previously rated by expert raters, and the scores provided by the raters in training are compared to the scores provided by the expert raters. In addition, the raters are also asked to provide rationale for their score assignment, which is compared to the rationale of the expert raters. During the practice scoring part of rater training, raters are given feedback regarding potential personal biases (e.g., personal feelings toward the question or personal feelings toward a given response) to reduce the construct-irrelevant variance due to these personal biases.

Qualifying Exam

Raters must pass a qualification exam to score responses in an operational setting. The objective of the qualifying exam is for raters to demonstrate that they can accurately and consistently provide scores across all score categories. The qualification exam often consists of a set of responses that aligns with each of the score categories. The set of responses and their alignment to the score categories are determined by experts and typically gathered in field testing or piloting of the item. Raters must achieve a certain level of agreement with the experts to pass the qualification exam (Van Moere, 2013). In essence, the goal of the qualifying exam is to ensure raters are calibrated to the rubric scale. Qualifying exams are often given to raters regularly to ensure that they continue to provide reliable and accurate scores. These exams act as a calibration and monitoring tool for raters and realign their scoring with the rubrics. The qualifying exam is a crucial component in providing a validity argument for the scores because it helps reduce the construct-irrelevant variance caused by raters.

Monitoring Operational Raters

Once a rater completes the training protocols, they are typically monitored in the operational setting through the use of blind validity checks and read-behinds to ensure raters are producing accurate and reliable scores (Engelhard, 2002). Blind read-behinds, or

backscoring, occur when operational responses are randomly selected and routed to expert judges/raters to ensure that the operational rater is providing reliable and consistent scores. If the operational scores differ from the expert raters' scores, the rater is removed from the operational setting and returned to the training protocol.

In addition to blind read-behinds, raters are often monitored over time through performance data and metrics. Validity responses is one monitoring method used at different points during the operational rating process (McClellan, 2010). Validity responses are responses that are scored by expert raters and then randomly seeded into an operational rater's queue. Since operational raters do not know the difference between a validity response and a regular response, the discrepancy between the score provided by the expert raters and operational raters is used as a method for continual monitoring of raters. Other metrics and data are also used to monitor raters, such as the score distributions of a rater and their agreement with other operation raters that score the same response (Engelhard, 2002; M. Graham et al., 2012). Raters can also be monitored through a re-examination of the qualifying exam at the beginning of each scoring session or at various specified time intervals.

Selecting Appropriate Raters

There are many considerations when hiring and selecting raters for particular items and assessments. Some criteria for selecting raters include educational background, subject and domain expertise/knowledge, and prior rating experience or teaching experience. The selection process of raters should also reflect the stakes of the exam. That is, the selection process should be more rigorous for exams with high stakes (Guo & Lei, 2020).

Research on Rater Training Protocols

Much of the early research on raters focused on the various aspects of the training procedure and how it affected score outcomes. Previous research indicates a few things for assessment systems to consider when developing a training protocol to lead to better future outcomes (Wolfe & McVay, 2012). It is important to demonstrate the existence of rater errors, such as leniency and severity of raters, and how to avoid these errors (Bernardin, 1978; Ivancevich, 1979; Uggerslev & Sulsky, 2008; Wind, 2019). Previous research also suggested that training protocols should directly precede operational scoring because it leads to more consistent and accurate scores (Attali, 2019; Roch & O'Sullivan, 2003; Sulsky & Day, 1994). Additionally, repeated practice scoring from the training protocols has been shown to improve the rating accuracy of operational raters (Hauenstein & McCusker, 2017). The attitudes of raters toward the training protocol have also been shown to affect the overall performance of raters (Noonan & Sulsky, 2001).

In addition to the above factors, the type of rater training method can influence rater effects. Two commonly used methods for rater training are the spiral and sequential methods. The spiral method trains raters on multiple items in a single training session.

The sequential method trains raters on a single item per training session, and every new item requires a separate training session. Moon and Hughes (2002) compared the spiral and sequential methods using a set of raters and found that, on average, the spiral method led to more accurate and reliable results from raters.

Measuring Rater Effects

Many psychometric models have been developed to measure the effects of raters, monitor raters over time, and classify raters. One of the first models to measure raters' effect from rater data was the rating scale model (Andrich, 1978). The rating scale model uses a Rasch measurement model with a facet for raters. The locations of the rater facets on the construct provide information regarding the severity of raters. Similar rater effects have also been measured through other latent trait models (e.g., Engelhard, 1994; Patz et al., 2002; Wolfe & McVay, 2012). The accuracy of operational raters can be modeled using an unfolding model known as the hyperbolic cosine accuracy model (J. Wang et al., 2016). The model uses accuracy ratings defined by the agreement between operational and expert raters to identify groups of raters that score similarly. In some cases, rater effects are measured over time by measuring the change in levels of accuracy of a rater or measuring how raters change their use of the score categories (e.g., Myford & Wolfe, 2009; Wolfe et al., 1999). These measurements are known as rater drift effects and are used to monitor operational raters throughout the scoring process.

MACHINE-BASED SCORING SYSTEMS

Machine-based scoring systems run the gamut from mark-sense readers to programs containing sophisticated scoring models of written or spoken work. None of these is “smart” in the sense that it can reason, problem-solve, create, and conclude. But most contain programming for relatively focused tasks that are reliable, modeled after expert (or human) behavior, and very efficient. For mark-sense readers, the programming is associated with determining the location of a “mark” (usually a bubble filled in with a No. 2 pencil) and associating it with a correct or incorrect response. It is sophisticated in that it can adjust for minor location deviations, can determine an intended mark from a stray one, and may be able to ascertain whether the pattern of marks is similar enough to another pattern in the batch to flag it for possible cheating. Algorithms that score essays compare a candidate essay against a model developed from a sample of essays that human raters have scored. The model may contain elements that evaluate grammar, organization and development, mechanics, syntax, and content. It may also be able to detect essays written in bad faith, ones that are plagiarized or off topic, or ones that present threats to self or others.

In this section, we will turn our attention to machine-based scoring applications and technology associated with tests, assessments, and surveys. Investment in this technology is designed to evaluate performance in a valid, reliable, and cost-effective

manner. Much has been accomplished, but there is a long way to go. The fourth edition of this volume reviewed the development of mark-sense readers, optical mark recognition, optical character recognition, and intelligent character recognition (Microsoft Corporation, 2022), and the reader is referred to that edition for more information on these long-standing devices. Consequently, we begin with a short review of more recent unique technologies that can assist in the delivery and data collection of performance assessments.

Data Capture

Many entry modes exist for written responses to an assessment. For computer-based assessments, the traditional mechanism is to type an answer on the keyboard. However, this requires that test takers have the ability to type. A tool for converting text into electronic forms is entry via tablets or touchscreen computers. In this case, a user would write the (block-written or cursive) text using a stylus on a screen that is simulated to look like paper or a test form. The text is then postprocessed and converted to electronic format. There are two advantages of using this approach over processing paper-based written text through intelligent character recognition. First, the response is converted directly to text rather than undergoing a separate scanning procedure to convert it to a scanned image for intelligent character recognition. Second, and importantly, the test taker can review and edit the results of the text conversion. Third, the design of the simulated paper corrects for skew or layout challenges. Most programs check for transcription errors through spellcheck, though sometimes errors can be introduced or inappropriately flagged by machines. For example, it is not uncommon for young children in the United States to write exclusively in block text (Loewus, 2016). Many of these individuals will write in all capitals. Spellcheck may flag words written this way as incorrect. Once converted, the text can be self-edited by the test taker, reducing errors that might be inadvertently introduced by postprocessing.

With speech recognition technology, dictated speech is converted into text as part of a transcription process or as part of the construct itself (e.g., English learner speaking assessment). This may be done as a testing accommodation or in generating a report that might, in its original form, be oral in nature (e.g., dictating the results of a dissection while the dissection is occurring). Alternatively, it may allow those who do not possess sophisticated keyboarding skills (e.g., young children) to express their performance orally but be processed or audited in written form. Many word processors support dictation (e.g., Microsoft Word), which has reasonable recognition accuracy rates (W. Xiong et al., 2018). The accuracy rate may be improved by “training” the processing engine to adapt to the speaker’s pronunciation and meter. The engine’s performance may also be enhanced by providing supplemental dictionaries. So, for instance, the dictation programs for radiology may have increased recognition rates by focusing on a more restrictive or specialized vocabulary (Ghai & Singh, 2012).

Data Integrity

The technologies described above come with varying degrees of accuracy and reliability. While long-standing devices such as mark-sense scanners are highly reliable, it is possible to introduce artifacts with even something as simple as this technology. The scanner may pick up stray marks made by the test taker, or the medium used for marking the answer sheet may be inappropriate (e.g., using a red pen instead of a blue or black pen). It is possible that the test taker changed their answer but did not erase the original answer thoroughly. Consequently, the machine picks up two responses instead of the intended one and has to be programmed for what to do in that situation. This can negatively impact the validity of score interpretation if the programming selects the unintended answer. The printing of the form itself may be skewed so that none of the answers can be accurately scanned. The best first step here is to visually inspect all the answer sheets to ensure that test takers conformed to the marking guidelines. A second check would involve a data validity check. Once the information has been processed, a follow-up can be made to establish that test takers marked their responses within an acceptable range. For example, the scan sheet may allow responses from a to f, but the test only had a response range from a to d. If a test taker marked alternative e, we would conclude that the individual made a mistake, but perhaps for a different reason than if they had selected an incorrect alternative in the range of a–d. A third check would be one that would address the possibility of cheating. In an earlier section of this chapter, we described security software that would estimate at a particular confidence level whether the answers had been copied.

Many data capture mechanisms perform a quality assurance check in the processing software. If the processing/scoring software encounters a character it does not recognize, it will flag it, display the original handwriting, and ask an operator how to proceed with processing the response (Microsoft Corporation, 2022). The optical character recognition operator will then supply the correction. It is possible to take samples of the same writing sequence and ask multiple operators to correct the same sequence of “flags” as a measure of agreement. Even then, those scoring the written responses may have to make a series of deliberate decisions about how best to modify (or not) the writing for scoring. For example, part of the algorithm that prompts the software to flag a response may be word misspellings if it took the most likely character representation as the correct one. Alternatively, the expected language for transcription might be English, but the writer included Spanish words or phrases. These words or phrases would likely be identified as errors unless a correction (or compensating algorithm) is in place. An important decision affecting the construct measured is whether spelling and punctuation errors are corrected or maintained during transcription or whether spelling and punctuation are even part of the construct. When handwriting is transcribed through a tablet, the test taker has the option of self-correcting words through immediate feedback. However, the question of how the writing might be modified to process it for scoring remains the same.

Processing oral (i.e., dictated) responses has its own data integrity challenges. The tolerance levels for accents differ depending on the software used, the spelling of the response is assumed to be correct, and each language has its quirks. For example, English requires the transcription software to determine the unique context of homonyms (e.g., to, too, two). In other languages, the same word can mean multiple things depending on the speaker's inflection. Moreover, it is important that punctuation be accurately accounted for so that the beginning of sentences is properly accounted for. The only way to determine the machine transcription or scoring accuracy is to validate it against a human rater or have the speaker review the transcription and make corrections. The software employed to transcribe or score responses from handwritten or oral responses can usually place a confidence estimate as to how likely the response was transcribed or scored accurately. These estimates can be validated against empirical evidence collected from human-scored responses for reliability and/or validity estimates. There is some evidence that children with learning disabilities benefit from dictating their responses rather than typing or writing them, especially for extended responses (MacArthur & Graham, 1987).

Interactive Computer-Based Systems

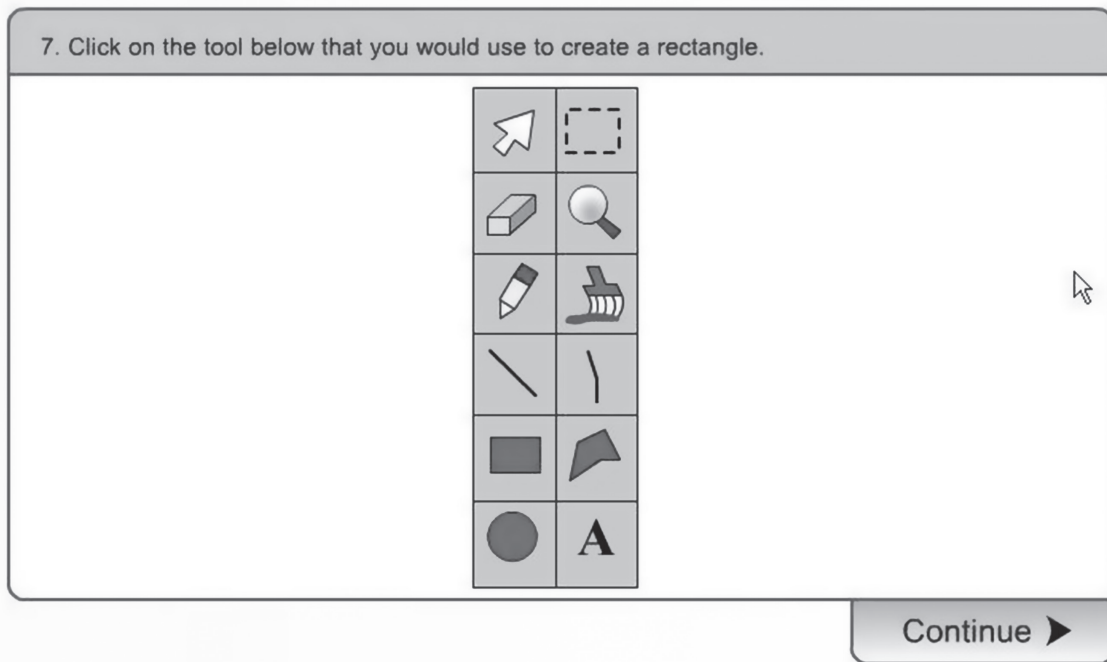
Computer-based innovative test items are designed to achieve a variety of outcomes. They may have a performance-based component that cannot be captured through a more traditional display of items (e.g., math items with formulas or graphing items), they may be designed to be more engaging or motivational, or they may be created to capture a repetitive behavior or action that could not otherwise be obtained (Parshall & Harmes, 2014). Scores may be validated using processes such as *rubric validation*, whereby content experts interact with the item to understand how it is scored or observe patterns of scoring in live data either randomly and/or in a targeted manner to determine how outliers are managed and typical responses are managed, etc. Scoring rules are defined by content experts and tested by them.

In a “hot spot” item, the test taker is asked to click on or touch the screen area that represents his/her response to a question or task. This is illustrated in Figure 8.1.

Figure 8.2 is an example of a “drag-and-drop” type item. It is similar to matching-type questions, except that it relies on a visual stimulus as the stem. Its main advantage is that test takers can visually confirm their choices, and the task is slightly more enjoyable than simply matching two columns of information.

Audio prompts can be used interactively with a computer in various settings. For example, audio prompts are employed in the College Board Advanced Placement Music Theory Exam, where the test taker is played a short music sequence and asked to transcribe it to a staff (College Board, 2021). Used in this way, the audio prompt played from a computer may facilitate the individualization of the process compared to playing the music to a group of test takers from an MP3 player.

A video prompt may use video imagery to convey prompt content in a way that would not adequately be described or addressed by words alone. In Figure 8.3, the student is

**FIGURE 8.1****An Example of a Hotspot Question**

Note. From "Improving the Quality of Innovative Item Types: Four Tasks for Design and Development," by C. G. Parshall and J. C. Harnes, 2014, *Journal of Applied Testing Technology*, 10(1), p. 7.

asked to imagine that they have been shrunk inside their classroom. You might elect to use a video-based stem if you believe that the imagery conveyed the situation or task more clearly. Or perhaps the modality would make the task more accessible if you thought that performing it would be hindered by an overreliance on reading ability.

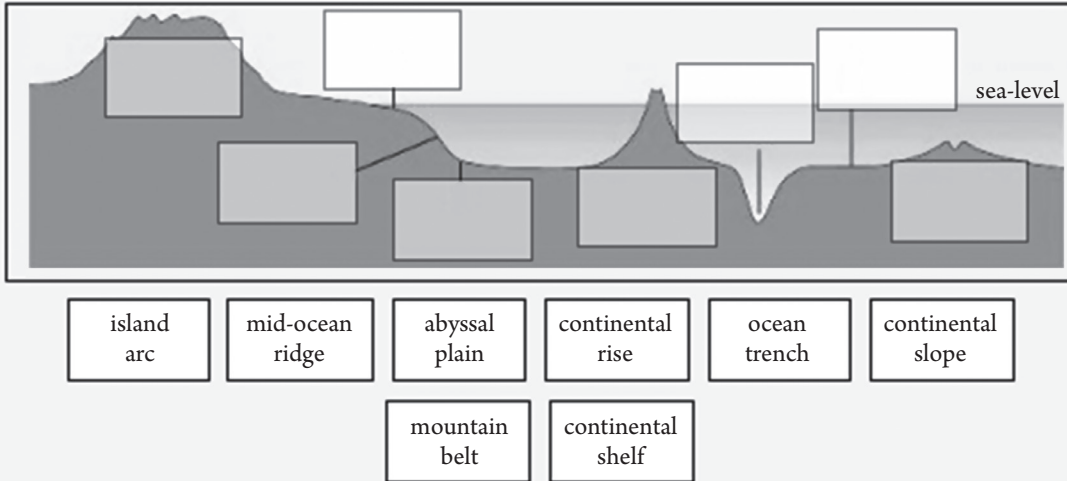
It could be argued that the text sufficiently defines the writing task asked of test takers as a writing prompt. However, the video prompt may provide additional context when the experience to be related is uncommon, new, or foreign. Even though the formats we have described have many compelling advantages, multiple-choice items are still the most common type of items administered on computer-based systems (Parshall et al., 2000).

Simulations as a Rehearsal or Demonstration of Skill

While the video prompt may serve as a setup for a performance assessment, it does not provide a responsive environment to demonstrate the desired or proxy skill. Simulations do. Simulation is the imitation of the operation of a real-world process or system over time (Banks, 1998). Simulations may be beneficial when exhibiting the actual skill is costly, dangerous, or may simply require much practice before it is acquired. Airline pilots will show that they can take appropriate steps in an airplane simulator during an emergency rather than employing an expensive aircraft containing live human beings.

Identify the features in this cross-section by dragging the labels into the boxes.

Use the mouse to drag the boxed words into the empty boxes. Alternatively, use the tab key to select an empty box, then use the space key to cycle through the options.



Check

FIGURE 8.2

An Example of a Drag-and-Drop Question

Note. From *Drag and Drop Onto Image*: Screenshot by Moodle, https://moodle.org/plugins/qtype_ddimageortext

In the education environment, science students may practice a frog dissection using simulation software before they perform the actual procedure. Knowing where to make an incision, how to identify and harvest the frog's organs, and techniques for moving from one part of the frog to another are skills that can be realistically simulated and assessed on software before beginning the actual work. The software practice can reduce mistakes and raise the probability of a successful authentic performance. Scoring a simulation can provide more realistic measurement, but only if carefully and properly conceived. Scoring is based on expert-driven rubrics that consider the consequences of making mistakes. Figure 8.4 shows a screenshot of a frog dissection simulation.

Serious-Based Games

As discussed earlier, a serious-based game achieves a learning goal through the demonstration of acquired or refined performance (Krath et al., 2021; Landers, 2014). The term *game* implies that the activity is fun, but the expression *serious* indicates that the purpose of the activity is educational or otherwise enlightening. Assessment data collected through serious games have some advantages over traditional tests in that one can document performance change over time, there are repeated samples of similar behavior, and feedback is immediate. The main drawback is that nonrelevant skills may impact performance. The ability to respond may require computer skills (e.g., ability to use a mouse or keyboard) that are irrelevant to the construct being assessed.

**BOX 1**

A TRANSCRIPT OF THE VIDEO PROMPT

Tell this story.

It's a Friday afternoon and you and two friends walk into the classroom to check on your science experiment. But when you touch the liquid, it bubbles over and instantly shrinks all three of you down to two inches, or about 5 centimeters.

Describe what you will do to survive the weekend.

Where will you find food and water?

How will you seek out shelter?

Will you create a structure? If so, what classroom materials will you use? Will you create your own miniature city?

What hazards do you need to avoid?

What challenges will you face as a tiny human?

What will you do when you're bored? Maybe ride a paper airplane? Create your own sport with string and tape and leftover cardboard?

Describe your adventures as a mini-student. Be sure to use vivid imagery. And consider all the ways you can be creative with your classroom setting.

FIGURE 8.3**A Screenshot of a Video Prompt**

Note. From *Writing Prompt: You Have Just Been Shrunk Down to Two Inches Tall* [Video], by J. Spencer, 2019, October 14, YouTube. <https://www.youtube.com/watch?v=Dzhxq3nCCqU>

**FIGURE 8.4****A Screenshot of Dissection Simulator Frog Edition**

Note. From *Dissection Simulator* [Screenshot], by Steam, 2023, April 11. https://store.steampowered.com/app/1046910/Dissection_Simulator_Frog_Edition/

There are five elements of a serious-based game (<https://www.game-learn.com/all-you-need-to-know-serious-games-game-based-learning-examples/>). These include the development of a story, a conversion of the situation to a game, the provision of immediate and individualized feedback, a simulation of engaging behavior, and a learning goal.

An example is the adventure game *80 Days*, as illustrated in Figure 8.5. Loosely modeled after the Jules Verne novel *Around the World in 80 Days*, the players learn geography by traveling around the world, having to plan their trip, making route adjustments as necessary, buying and selling materials in different markets, making health decisions, and achieving their goal in “80 days” or fewer. The degree to which gamers follow the rules of the game and make good decisions is the basis for scoring. “Goodness” may be based on expert-driven rubrics or on some other basis (e.g., efficiency).

ARTIFICIAL INTELLIGENCE SCORING

Artificial intelligence (AI) scoring is the ability of computer technology to evaluate and make score predictions for some type of open-ended performance. The most common machine scoring assessment is for written prose (Shermis et al., 2013), but it has been expanded to short-form constructed responses and speech. The examples in this section are based on these three response modalities. The scoring technology to accomplish this has come under several different names, including its original incarnation, automated essay grading, and one that has been in place since the early 2000s automated essay scoring (AES). It is also referred to as automated essay evaluation or automated writing evaluation. The latter terminology suggests that the technology

**FIGURE 8.5**

A Screenshot From the Serious-Based Game *80 Days*

Note. *80 Days* [Screenshot]. Copyright 2014 by inkle. Reprinted with permission.

can do more than just score essays. For example, several systems are supplemented with natural language processing (NLP) capabilities that can analyze the structure of writing and provide qualitative feedback (e.g., “Your essay appears to be about the role of media in a democracy”) that may or may not be a component of the scoring process. Most of the initial applications have been in English, but development has been reported in multiple languages, including, for example, Japanese (Ejima & Takeuchi, 2022; Ishioka & Kameda, 2006), Chinese (Chang & Lee, 2009; Qian et al., 2020), and Swedish (Lilja, 2018; Östling et al., 2013). (See also Table 8.2 for a more comprehensive list of non-English AES scoring engines.) All but the most enthusiastic proponents of AES suspect that there are forms of writing that will always be difficult to evaluate (e.g., poetry). However, the technology has been applied to the score predictions of multiple genres and populations at various writing levels with some degree of success (Shermis, 2014). It is important to note that machine scoring models may employ some aspects of artificial intelligence in their algorithms or they may be based on static data models (Mayfield & Black, 2020).

In the United States, the technology was spurred on by the initial adoption of the Common Core State Standards (Porter et al., 2011), where one of the goals was to develop assessments that would address higher order thinking skills and encompass more writing. While most states eventually walked back their commitments to the Common Core State Standards and the consortia that promoted them (i.e., Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balanced), many states began to incorporate the general principles of the Common Core in their state standards. This meant that the amount of writing assessed increased and the need

to come up with an economical way of evaluating large volumes of writing came into demand. The first widespread demonstration of the AES technology began with the Hewlett trials of 2012. The Hewlett trials are described in a later section; they marked a new wave of development for AES technologies.

For the most part, AES models are based on human ratings of a sample of essays either generated from a specific prompt (e.g., “prompt-specific”) or accumulated across essays within a particular writing genre and grade (e.g., “generic”). To better understand this process, reviewing an example of a human writing model is helpful.

Rubrics

One of the challenges in creating AI models is that rubrics are often hard to define, vague, and incomplete. In this context, AI refers to the ability of a computer to emulate intelligent human behavior (Simon, 2019). A search for a conceptual definition of “good writing,” for example, is broad and imprecise, and those who assess in that domain end up using operational definitions grounded in rubrics. For the most part, the rubrics are based on characteristics of writing (traits) or components of writing (i.e., analytic scoring; Stiggins & Bridgeford, 1983). One trait rubric is the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2002). The CEFR was formally introduced in 2002 as a collective project of the European Union to establish criteria and assessment procedures for evaluating individual language performance. The CEFR describes language proficiency in reading, writing, speaking, and listening on a six-level scale. It is used for instruction and assessment to create a system that has similar meaning across all languages covered by the European Union. The CEFR is just one popular conceptual framework; there are many others with their associated rubrics (Shermis, 2022).

AES

In these next sections, we discuss products from ETS for automated essay and speech scoring. They were chosen because they are well documented, used in a number of different applications, and highly researched. Table 8.1 lists other English AES systems that are available for use and may incorporate other features. e-rater is an automated essay evaluation by ETS (Attali & Burstein, 2006; Burstein et al., 1998; Burstein, Tetreault, & Chodorow, 2013). The e-rater engine first became operational in 1999, when it was deployed to provide one of two scores for essays on the writing section of the Graduate Management Admissions Test, a high-stakes, large-scale assessment. In conjunction with human ratings, e-rater is also currently used for the computer-based Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) and as a check score (i.e., score from a second rater) for the GRE. In the Criterion Online Writing Evaluation service—a web-based writing tool that helps students plan, write, and revise their essays—e-rater is used to foster the best instructional and assessment practices (Burstein, Tetreault, & Madnani, 2013).

Table 8.1 Comparison of English-Based Automated Essay Scoring (AES) Systems

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training	Language
AutoScore	American Institutes for Research (now Cambium Assessment)	Natural language processing (NLP)	Style and content	N/A	2,000–3,000	English
Bookette	CTB McGraw–Hill	NLP	Style and content	Writing Roadmap	250–500	English
CRASE	Pacific Metrics (now ACT) (Lottridge et al., 2013)	NLP	Style and content	N/A	100 per score point	English
e-rater	ETS (Burstein et al., 1998)	NLP	Style and content	Criterion	100–1,000	English
Intelligent Essay Assessor	Pearson Knowledge Technologies (Landauer et al., 1997)	LSA	Content	N/A	100–300	English
Intellimetric	Vantage Learning (Elliot, 2003)	NLP	Style and content	MY Access!	300	English Chinese Hebrew Bahasa Malay
Lexile Writing Analyzer	MetaMetrics (Burdick et al., 2011)	NLP	Style and content	N/A	None—uses a fixed model	English
LightSIDE	Teledia Laboratory (Mayfield & Rosé, 2013)	Statistical	Content	N/A	300	English

Note. Adapted from “State-of-the-art Automated Essay Scoring: Competition, Results, and Future Directions from a United States Demonstration,” by M. D. Sheremis, 2014, *Assessing Writing*, 20, 53–76, p. 63 (<https://doi.org/10.1016/j.asw.2013.04.001>).

One of a family of scoring engines used at ETS, e-rater aligns a defined writing construct with NLP methods to identify linguistic features in student and test-taker writing for scoring and evaluation (e.g., diagnostic feedback). Consideration of aspects of the writing construct in earlier system development (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013) and more recent research and development support the enhancement of the system’s construct coverage about the structure of argumentation (Beigman-Klebanov et al., 2013), discourse coherence (Burstein, Tetreault, & Chodorow, 2013), and vocabulary usage (Beigman-Klebanov et al., 2013). The e-rater software currently identifies and extracts several feature classes for model building and essay scoring using statistical and rule-based NLP methods. These feature

sets are proxies for the more conceptual attributes of a human-based writing construct (Attali & Burstein, 2006).

Feature development and re-evaluation are dynamic, and specific feature variables may vary as the system is updated with new releases. Individual feature classes represent an aggregate of multiple features. The feature classes—the variables of writing that constitute the construct model developed for score prediction—include the following: (a) grammatical errors (e.g., subject–verb agreement errors), (b) word usage errors (e.g., *their* versus *there*), (c) errors in writing mechanics (e.g., spelling), (d) presence of essay-based discourse elements (e.g., thesis statement, main points, supporting details, and conclusions), (e) development of essay-based discourse elements, (f) style weaknesses (e.g., overly repetitious use of vocabulary), (g) two content vector analysis-based (Wolfe & McVay, 2012) features to evaluate topical word usage, and (h) a feature that considers correct usage of prepositions and collocations (e.g., *powerful computer* vs. *strong computer*; Futagi et al., 2008) and sentence variety. The set of features in (h) represents positive features, rather than errors in conventions. Because proper usage of English prepositions and collocations is especially difficult for English learners, the addition of these features also expands e-rater’s ability to recognize characteristics of writing that are important for assessing nonnative writers. Figure 8.6 illustrates the breakdown of the e-rater model. More details about specific features aggregated within a feature class may be found in Attali and Burstein (2006).

Human-assigned scores are used to build e-rater models and are the basis for building models for most of the other automated scoring engines. A randomly selected training sample of at least 250 human-scored essays is processed through e-rater, extracting the features described above. Features are aggregated into conceptually related groups and converted to a vector (Downey et al., 2008) of numerical feature values. Using a regression modeling approach, the values from this sample are used to determine an appropriate weight for each feature (Attali & Burstein, 2006). The same process is performed vis-à-vis feature extraction and conversion of features to a vector format to score a new, unseen essay during a test administration. To compute the final score prediction, these values are then multiplied by the weights associated with each feature, and a sum of the weighted feature values is calculated (Attali et al., 2010). In addition to providing holistic scores using construct-aligned language features, e-rater provides advisories for detecting such construct-irrelevant elements and attempts to enhance the overall score (Higgins et al., 2006). Many features used for e-rater scoring are also used for feedback in Criterion, a portfolio program used for writing instruction. These include the grammar, usage, mechanics, and styles features and the organization and development features described above.

More recently, other scoring engines have incorporated additional feature sets to include predictors such as n-gram vectors (i.e., content), latent Dirichlet allocation (i.e., topic modeling), deep neural network embeddings, Bidirectional Encoder Representations from Transformers (BERT), and passage similarity metrics as ways to improve outcomes (Lottridge et al., 2020a).

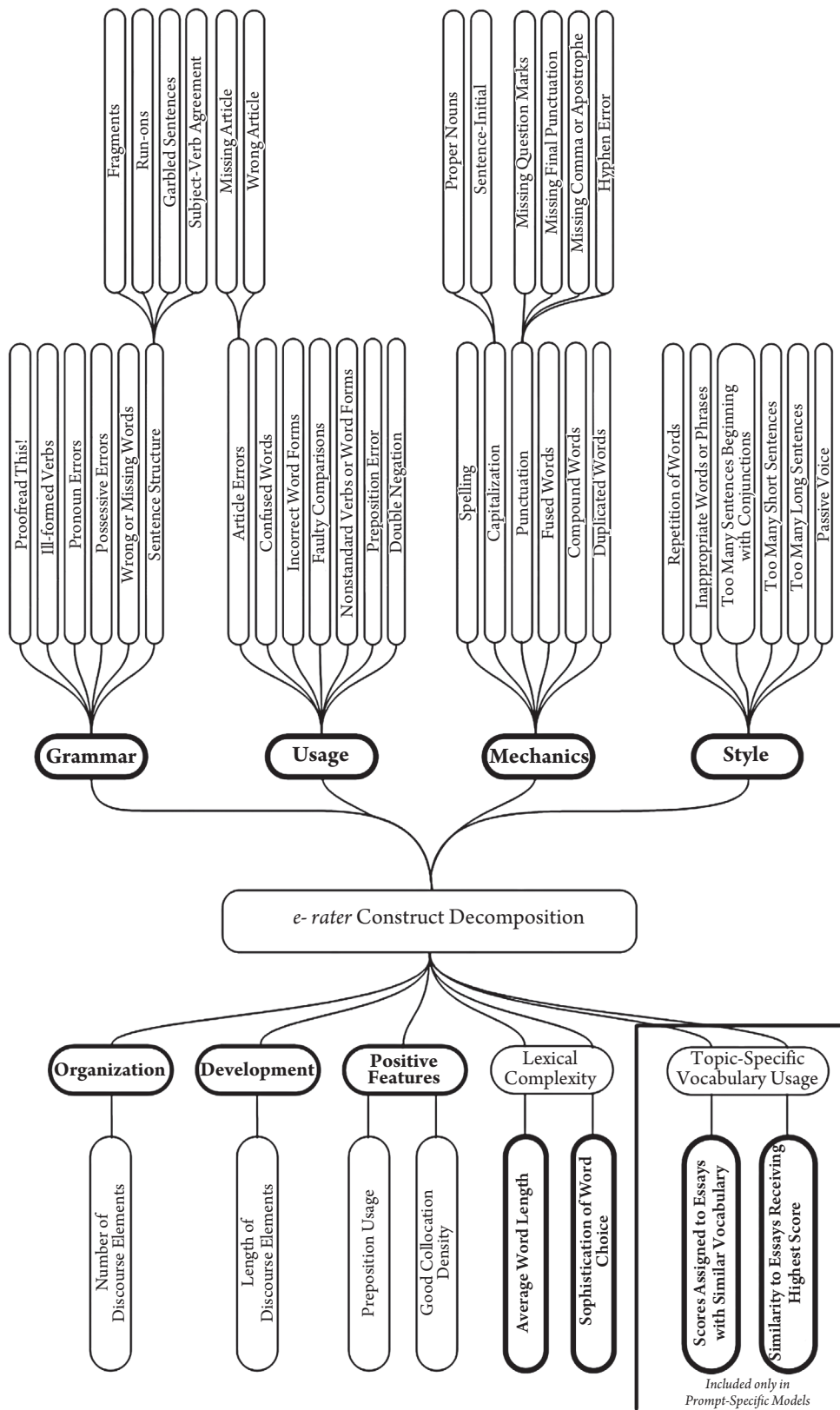


FIGURE 8.6

Organization and Construct Coverage of e-rater Version 19.1, Released in 2019

Short-Answer Scoring

With AES, the models are formulated through an assemblage of features and the process then compares a candidate essay against the model and predicts a score. The mechanical aspects of writing are often expressed as a ratio of the propensity of mechanical errors to the volume of writing, while the content may be expressed as “distance” between the semantic dimensions of the model to the semantic dimensions of the candidate essay.

With short-answer scoring, the focus tends to be more on evaluating the content of what was written rather than the response’s style. Two different approaches are taken for model building. In the older approach, answers, answer variants, answer ranges, and sometimes wrong answers are programmed and compared against a candidate response. The closer the match, the higher the score. Answer variants are sometimes linked to a corpus of synonyms so that the word “ruler” or “monarch” might be considered (or not) equivalent to the word “king.” If answers are modeled on human responses, it is usually helpful to establish the range of responses or test out possible exceptions to the programming. Some scoring engines have routines to help evaluate the candidate answers in the context of using passive voice (e.g., “Carthage was destroyed by Rome” versus “Rome destroyed Carthage”; Leacock & Chodorow, 2003).

A more recent approach employs the multiple feature extractions mentioned earlier, including latent semantic analysis and n-gram vectors for content, latent Dirichlet allocation for topic modeling, deep neural networks and machine learning, BERT, passage similarity metrics, and a variety of linguistic variables. These features are then compared against a rubric and a score is predicted. Figure 8.12, which illustrates the process graphically, is referenced later in the section on the validity of machine scoring. For more modern scoring algorithms, the process for scoring short-answer questions and more extended essays is essentially the same. Topic modeling, for content, is addressed in the next section of this chapter (Ormerod et al., 2022).

The National Assessment of Educational Progress (NAEP) program recently ran a competition on the automated scoring of short-form constructed responses for NAEP Reading in Grades 4 and 8. Twelve competitors were provided item materials, a randomly selected training and validation set, and asked to make predictions for a test set where only the text of the response was provided using the NAEP 2017 Reading data. In these trials, human rater performance was $\kappa_w = .91$; for the competition winners, the scoring performance was $\kappa_w = .89, .88$, and $.87$. Competitors were also asked to submit reports on the fairness and transparency of their models (Shermis, 2024).

Automated Speaking Assessment

Automated scoring of speech grew out of the need to address the volume of second-language (L2) learners trying to acquire English-speaking skills (Howson, 2013; Zechner & Loukina, 2020). As discussed earlier, the European Union uses English as the official language of commerce and employs an assessment framework that can rely on both human and machine scoring of speech tasks (CEFR; Council of Europe, 2001; Green, 2013).¹

Most people have had casual interaction with an automated speech recognition (ASR) system either through digital assistants (e.g., Alexa, Siri, Google Assistant) or by dictating into a speech-to-text translator program. These applications are different from speech scoring systems in that they are optimized for native speakers or can be adapted through training and software settings (i.e., to account for dialects). Even with highly sophisticated ASR systems, there are subtle indications that they may be less than perfect. For example, you may ask a digital assistant to turn *out* the lights when the programmed protocol is to turn *off* the lights. Alternatively, you may dictate, “The rain in Spain falls mainly on the *plane*” when you meant “The rain in Spain falls mainly on the *plain*.”

Automated speaking assessment follows a paradigm similar to that of AES. Language- and speech-related features are extracted, and a scoring model is used to compute a score based on a combination of these features. The main difference between text and speech is that the word identities are unknown in speech and an additional component is needed—the ASR system—to generate a word hypothesis from the test taker’s digitized spoken response to an item prompt. This process is illustrated in Figure 8.7. In addition to L2 assessment, automated speech scoring is used in related applications such as the automated assessment of job interview performances (L. Chen et al., 2016) and automated evaluation of oral presentations (Ramanarayanan et al., 2015).

Word error rate (WER) is a standard metric for measuring speech-to-text accuracy of ASR systems.² WER uses the Levenshtein distance to determine the fewest number of steps needed to convert the engine-transcribed text to the target (human-transcribed) text. For native adult speakers, WER is relatively low. For example, Microsoft claims its ASR system has a word error rate of 5.1%. Google reports a WER of 4.9%. For comparison, human transcriptionists average a word error rate of 4% (H. Chen, 2021). However, WERs for spontaneous nonnative speech are substantially higher, on the order of 20%–30% or more. Factors that affect recognition accuracy include the variation in native languages of the test takers, variations in their speaking proficiency, suboptimal recording conditions (e.g., environmental noise, background talk), and speaker idiosyncrasies (some speakers are inherently harder to recognize by ASR systems than others; Hsu et al., 2021; Shermis, Burstein, et al., 2015; C. Wang et al., 2021).

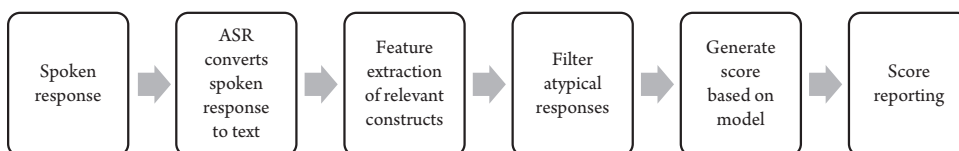


FIGURE 8.7
Standard Process Approach for Automated Speech Scoring

Note. Adapted from “Overview of Automated Speech Scoring,” by K. Evanini & K. Zechner, in K. Zechner & K. Evanini (Eds.), *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech* (pp. 3–20), 2020, Routledge.

ASR is typically trained using machine learning algorithms that are based on large amounts of speech data. The process of training an ASR system involves several steps, including:

1. **Collecting speech data:** The first step in training an ASR system is to collect a large amount of speech data. These data can come from a variety of sources, such as audio recordings of human speech, transcribed audio recordings, or text-to-speech data sets.
2. **Preprocessing the data:** Once the set of data has been collected, it must be preprocessed to prepare it for use in training the ASR system. This can involve tasks such as converting audio files to a standard format, removing background noise, and segmenting the audio into smaller chunks.
3. **Feature extraction:** The next step is to extract features from the speech data. These features may include things like pitch, duration, and frequency of the sound waves in the audio. The goal is to extract features that are relevant to the task of recognizing speech.
4. **Training the model:** With the preprocessed and feature-extracted data, the ASR model can be trained using machine learning algorithms such as deep neural networks. During training, the model is exposed to large amounts of labeled speech data, and it learns to recognize patterns in the data that correspond to different words and phrases.
5. **Evaluation:** After the model has been trained, it is evaluated using a separate data set of speech data. This evaluation is used to determine the accuracy of the model and identify areas where it may need to be improved.
6. **Fine-tuning and optimization:** Based on the results of the evaluation, the ASR model may be fine-tuned and optimized to improve its accuracy. This may involve adjusting various parameters of the model or collecting additional data to retrain the model.

Overall, the process of training an ASR system is iterative, and it may take many rounds of training, evaluation, and fine-tuning to develop an accurate and reliable system (Evanini & Zechner, 2020).

SpeechRater

SpeechRater is a product of ETS that was deployed to score the speaking section of TOEFL Practice Online in 2006. SpeechRater is used here as an example, but there are many commercial systems available (Cheng et al., 2014; Pearson, 2017). This system helps students prepare for the TOEFL iBT. Unlike most other speech scoring systems, SpeechRater is intended to provide scoring for assessments that elicit a wide range of evidence of speaking proficiency (i.e., not only pronunciation) and to elicit spontaneous and natural speech from the test candidates as opposed to mere reading or repetition (Zechner et al., 2009).

The tasks scored by SpeechRater are modeled on those used in the speaking section of the TOEFL iBT. These tasks ask the test taker to provide information or opinions on familiar topics based on their personal experience or background knowledge and respond to reading or audio stimuli related to campus life and academic situations, such as lectures. The speaking time per item is about 1 minute. Test takers are scored on a scale of 1 (lowest proficiency) to 4 (highest proficiency), with a score of zero assigned to responses that do not address the task.

The recorded spoken response of the test taker is first sent to an ASR system, which generates a hypothesis, a string of words along with timestamps, that provides its best estimate of what the speaker said. Next, several feature extraction programs generate a vector of features indicative of various aspects of speech, such as fluency, pronunciation, prosody, vocabulary, and grammar, as exemplified in the TOEFL rubric (ETS, 2013). Finally, the features are combined in a scoring model to predict a score for each spoken response. The six-item scores of the test taker are then aggregated and scaled and placed in a score report for the test taker, along with information about how to interpret the score.

The TOEFL rubric specifies three dimensions of attributes that contribute to the score of a response:

- delivery (low-level technical features of speech production—e.g., pronunciation and fluency),
- language use (formal cues of linguistic competence—e.g., grammar, diction), and
- topic development (higher level semantic, pragmatic, and organizational aspects of the response).

Delivery features can be extracted most easily directly from the output of the speech recognition system, while language use is more difficult to address, given the constraints of recognition accuracy. The appropriate development of the topic is even more challenging to assess without an accurate response transcript.

Most of the features used in the early versions of SpeechRater addressed the delivery aspect of the TOEFL speaking construct (i.e., fluency and pronunciation). More recently, extended sets of features were developed and added to SpeechRater, including features measuring aspects of rhythm (L. Chen & Zechner, 2011), vocabulary (Yoon et al., 2012), grammar (Bhat & Yoon, 2015; Chen & Zechner, 2011), content (Xie et al., 2012), and discourse (X. Wang et al., 2016). The statistical model SpeechRater uses to predict the score based on these features is a multiple linear regression model. However, promising experiments have been performed using decision trees as well (Xi et al., 2008).

Like the advisories generated by e-rater, SpeechRater has a so-called filtering model that can identify responses that should not be scored automatically because of certain deficiencies or anomalies (Higgins et al., 2011; Yoon et al., 2011). Compared to

written texts, such nonscorable responses comprise a much more comprehensive range of phenomena—for example, low audio quality due to technical issues in the recording process, background noise, background talk by other test takers, or non-English responses. The reliability and validity of SpeechRater is discussed at the end of the chapter.

AI Scoring Conclusion

Since 2004, the progress in automated speech evaluation methods and systems has been substantial, and state-of-the-art speech scoring systems can measure certain aspects of speech proficiency quite reliably, particularly when the spoken test responses are somewhat constrained and predictable. For more spontaneous and open-ended responses, progress has been slower, and automated speech assessment systems are currently unable to evaluate all the aspects of speaking proficiency that human expert raters are trained to consider and evaluate. This is largely because of limitations in current speech recognition capabilities when applied to nonnative speech by speakers with a broad range of speaking proficiencies.

Some areas for future work include the conceptualization and development of features in the domains of the speaking construct that have not yet been adequately addressed, improvements to the speech recognition system, and research into more challenging but more natural item types, such as dialogues.

TOPIC MODELING OF CONSTRUCTED RESPONSE ITEMS

As briefly mentioned in the previous section, various topic models are used to extract features that are then used as covariates in many machine learning algorithms for scoring short-answer items and essays. Topic models are mathematical and statistical models used to analyze sets of documents (i.e., corpora) and were initially developed as an indexing technique to summarize, compare, and classify large sets of textual data. The primary objective of topic models is to discover topics and the usage of topics within a corpus (Chemudugunta, 2010). Topics are textual features that occur across all documents and can often be thought of as latent clusters of words. The usage of topics is a textual feature that occurs within an individual document and expresses the use of topics within the document.

Although the use of topic models was initially developed outside educational measurement, they have been shown to be useful methods for a variety of applications related to scoring and test security, such as feature extraction for automated scoring engines, improving ability estimation under an item response theory framework, discovering topics that are elicited by different items, identifying misconceptions, detecting complex forms of cheating, and providing feedback to students (e.g., Foltz et al., 1999; Gierl et al., 2020; Mozgovoy et al., 2010; Wheeler et al., 2022; J. Xiong & Li, 2023). Many applications of topic models in measurement utilize the latent seman-

tic analysis (Deerwester et al., 1990) and latent Dirichlet allocation (Blei et al., 2003) models and their variants.

Latent Semantic Analysis

Latent semantic analysis (LSA) is one of the earliest topic models and was originally designed for indexing and comparing text from a corpus. LSA is similar to factor analysis and principal component analysis in that it extracts latent components (i.e., topics) from a corpus and detects the relationship between each document and each word with respect to those components (Landauer et al., 2013). Applying LSA requires a matrix representation of the corpus, which is referred to as the document-by-word matrix where each row belongs to a distinct document in the corpus, each column belongs to a word from the vocabulary (i.e., the set of all unique words that appear in the corpus), and the cells contain the occurrence of each word in each document. Figure 8.8 displays a document-by-word matrix for a corpus that contains D documents and W unique words.

The document-by-word matrices are typically sparse (i.e., they contain many zeros) because most documents use only a small subset of words from the vocabulary. This tends to cause meaningless comparisons between documents directly using the document-by-word matrix. LSA, therefore, applies singular value decomposition (Golub & Reinsch, 1971), a linear algebra technique, to the document-by-word matrix. Singular value decomposition calculates the observed variation across documents with respect to latent components by decomposing the document-by-word matrix into three orthonormal (orthogonal and standardized) matrices: a document matrix, a topic matrix, and a word matrix. The document matrix contains the relationship between each document and the latent components, the word matrix contains the relationship between each word and the latent components, and the topic matrix contains the relative weights of each latent component. Multiplying the three matrices results in the original document-by-word matrix, as shown in Figure 8.9.

The three orthonormal matrices constitute the underlying semantic space of the corpus. The semantic space of a corpus is defined as the mathematical representation for the meaning of words and documents that is captured by the co-occurrence of words

	Word 1	Word 2	...	Word W
Document 1	0	1	...	2
Document 2	0	0	...	0
⋮	⋮	⋮	⋮	⋮
Document D	2	0	...	0

FIGURE 8.8
An Example of a Document-by-Word Matrix

across documents. The objective of LSA is to reduce the underlying semantic space by removing components with relatively small weights from the document, word, and topic matrices (Deerwester et al., 1990). That is, the components that explain the least amount of variation across the documents are removed from the semantic space (Dumais, 1994). Selecting an appropriate number of components to remove depends on a variety of factors, including the number of documents, the number of words, and the intended use (see Wild, 2016, chap. 7).

After the components are removed, the matrices can be multiplied together to obtain a reduced document-by-word matrix, as shown in Figure 8.10. The reduced document-by-word matrix approximates the original document-by-word matrix, meaning the cells of the matrix now contain approximate values for the occurrence of words in each document. Additionally, the reduced matrix is no longer sparse (i.e., it now contains only a few zeros). This provides a richer representation of the corpus, which allows for more meaningful comparisons between documents (Landauer et al., 1998). Calculating the cosine similarity between rows of the reduced document-by-word matrix yields a useful measure of textual coherence between the documents (Foltz et al., 1998).

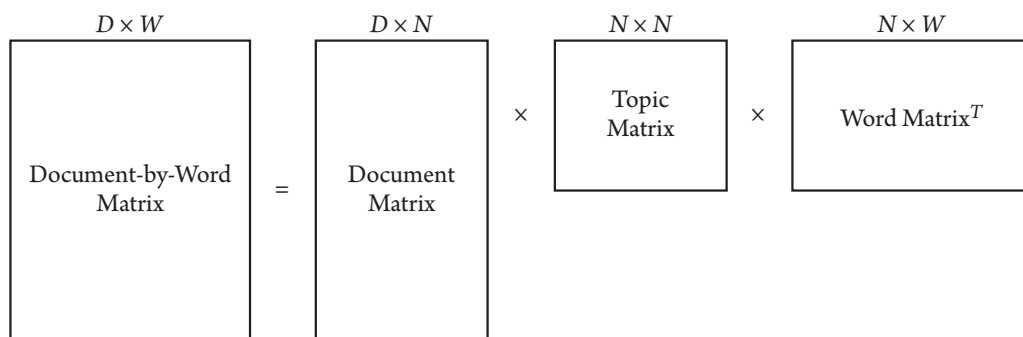


FIGURE 8.9
Illustration of Singular Value Decomposition on a Document-by-Word Matrix

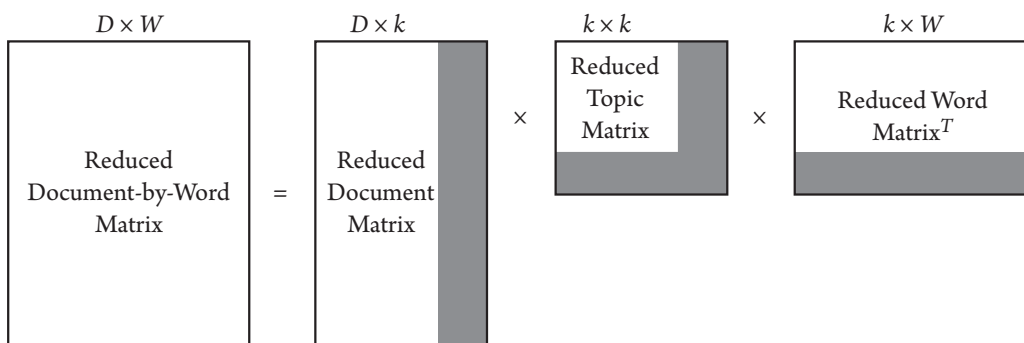


FIGURE 8.10
Illustration of Latent Semantic Analysis Dimensionality Reduction

LSA is used in a variety of ways for AES, since the features and topics extracted from LSA essentially reflect the language patterns used by students when responding to an item. For example, the document and topic matrices extracted from singular value decomposition can be used as a set of input variables for scoring algorithms. The similarity measures between documents can also be used to create textual features for scoring algorithms. Coh-Metrix (McNamara et al., 2014), a computer program for analyzing and extracting features from a set of documents, utilizes LSA to produce various metrics for written and spoken texts that are then used as covariates in scoring algorithms. Although LSA has been successful in providing features for automated scoring algorithms, applying LSA to unseen documents containing words that were not in the original corpus used to fit the model can be difficult and can cause the algorithms to underperform.

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a probabilistic mixed membership model used to analyze sets of documents (i.e., the corpus). LDA aims to estimate three sets of parameters: topics, topic proportions, and topic assignments. Topics are latent clusters of words and are defined as a probability distribution over all the unique words within the corpus (i.e., vocabulary). The topic distributions are corpus-wide features that express the probability of each word in the vocabulary occurring within the given topic. Topics identify common themes seen across all documents in a corpus. Topic proportions are latent clusters of topics and are defined as a probability distribution over all the topics. The topic proportions are document-wide features that express the usage of each topic within a document. Topic assignments are classifications that assign each word in a document to one of the topics (Blei et al., 2010).

Figure 8.11 displays a graphical representation of LDA. The graphical model has three rectangles, or plates, representing the textual data's replications or levels. Documents are considered replications since there are multiple documents within a corpus. Words are considered replications nested inside documents because documents contain multiple words. Topics are considered replications since there are multiple topics within a corpus. The circles represent the variables of the LDA model, and their placement within the plates shows the level at which the parameter occurs. Shaded circles

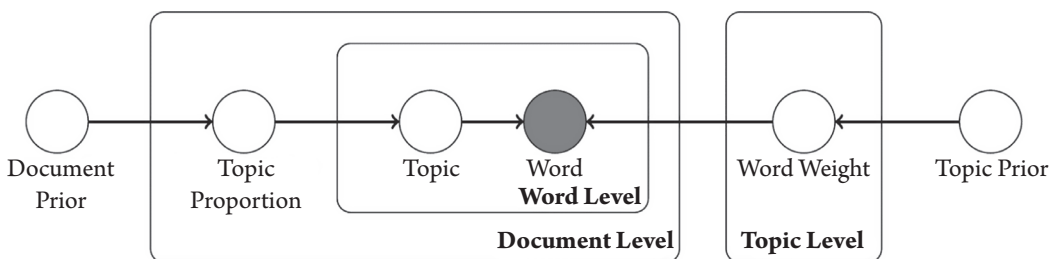


FIGURE 8.11
Graphical Plate Representation of Latent Dirichlet Allocation

indicate observed variables, and unshaded circles indicate latent variables. The topic proportions are located at the document level because each document is given a set of topic proportions that represent its use of each topic. The topic assignments are located at the word level because each word is given an assignment to a topic. The word probabilities are situated at the topic level because each topic is a probability distribution over all the words. The arrows represent the directional relationship between each variable. The variable at the head of the arrow (i.e., where the arrow ends) depends on the variable at the tail of the arrow (i.e., where the arrow originates).

Assuming the number of topics a priori, the LDA model assumes a generative process for constructing each document within a corpus. That is, given the number of topics, LDA assumes a document is generated using the following process (Blei et al., 2003):

1. Select the number of words from a Poisson distribution.
2. Select the topic proportions from a Dirichlet distribution.
3. For each word in the document:
 - a. Select a topic assignment, given the topic proportion, from a multinomial distribution.
 - b. Generate a word, given the topic assignment and topics, from a multinomial distribution.

This process is used to develop estimation algorithms to infer the topic, topic proportion, and topic assignment parameters of the model from the observed words.

Applying LDA to a set of documents is an exploratory approach because the number of latent topics is assumed before the model is estimated. Therefore, selecting an appropriate LDA model requires researchers to fit multiple candidate models using a different number of topics and then compare each model based on statistical fit indices and interpretability. Additionally, when LDA is used as a feature extraction technique for automated scoring, the number of topics can be selected based on various performance metrics. The number of topics that provides the most parsimonious model is usually chosen as the best-fitting model (Mardones-Segovia et al., 2023).

Considerations for the Application of Topic Models

Before fitting a topic model and extracting features from textual data, the documents must go through data cleaning steps (Boyd-Graber et al., 2014). First, all punctuation is removed and all words are converted to lowercase, ensuring that the model is not counting the same word twice because of capitalization. For example, the word *dunes* and *Dunes* should only count as one unique word rather than two. Next, the words go through a stemming process that involves retrieving the base meaning of a word from its different tenses. For example, the verbs *swim*, *swims*, *swimming*, and *swam* all carry similar meaning and might be converted to a single word, *swim*. This process also involves nouns; for example, *teen*, *teens*, *teenager*, and *teenagers* all have a similar meaning and could be converted to the word *teen*. The stemming process is intended to increase the clarity of the topics by using a single unique word. In this way, it enables

the model to cluster words with similar meanings together, thereby enhancing the interpretability of the topics (Schofield et al., 2017). The process of stemming also involves changing misspelled words into their correct forms. This process is context dependent, meaning there is no universal list of stemming words, and it is often tailored to the corpus.

The last data cleaning step is the removal of stopwords. Stopwords are common words that carry little information about the topics (Wilbur & Sirotkin, 1992). Stopwords tend to occur with high frequencies, and if they are not removed, they will tend to dominate the latent topics, reducing the interpretability of the topics (Choi et al., 2017). Generally, stopwords lists include words such as *the*, *a*, *are*, and *they*. There is no theoretically grounded approach to choosing and removing stopwords. Many techniques, however, aid in selecting stopwords, such as custom rules (Schofield et al., 2017), which are developed by researchers for their specific application. These allow the researcher to create a unique stopwords list specific to the intended analysis of the corpus. The TF-IDF index is another method for selecting stopwords (Robertson, 2004). This index uses term frequencies to decide whether a word is being used too often. Finally, the general stopwords rule (Manning et al., 2008) combines statistical approaches and customization to define stopwords. Previous research suggests customized stopword lists are needed for analysis of extended-response items (Choi et al., 2017).

Applications of Topic Models in Educational Measurement

Topic models are widely known in educational measurement as a feature extraction technique for automated scoring algorithms. For example, commonly used machine scoring programs, such as the Intelligent Essay Assessor (Foltz et al., 1998), Coh-Metrix, and Autoscore (Lottridge et al., 2020b), use LSA to extract meaningful document features. LDA has also been used to assess the viability of using topic proportions as covariates in a scoring algorithm. However, results showed that LDA methods performed worse than LSA methods (Kakkonen et al., 2006; H. Xiong et al., 2019).

The utility of these models, however, lies beyond automated scoring and in analyzing test-taker text for themes. For example, Shin et al. (2019) used LDA to analyze written responses to test questions to identify common misconceptions that could be used to develop distractors for multiple-choice items. Wheeler et al. (2024) demonstrated that LDA could be used to extract and summarize information about test takers' reasoning when responding to constructed response items. The results suggested that LDA can be used to provide validity evidence for test developers and provide educators with information to inform instructional approaches for essays. Duong et al. (2019) used LDA to analyze answers to an essay test requiring integrative borrowing. Responses of students taught to use instructional conversation (a form of collaborative learning) were compared to responses of students who had not. Results indicated that answers of students taught to use instructional conversation expanded beyond the characters and settings in the prompt and expressed an understanding of multiple perspectives compared to the answers of students who had not had an instructional conversation training. Students

who had not had instructional conversation training simply restated the prompt, did not change the setting, and only took a single perspective into account in their answers.

Choi et al. (2017) used LDA to detect latent topics to aid in identifying the semantic and thematic structure of students' writing. Topics that differentiated the types of words and the latent themes in good versus poor writing were detected. Kim et al. (2017) found that two of the three topics in the topic model of answers to a test of science inquiry knowledge corresponded to academic language and discipline-specific language as reflected in the use of scientific words such as *measure*, *effect*, and *hypothesis*. Answers making use primarily of the third topic used mainly everyday language. Kwak et al. (2017) reported a change in the use of topics detected by LDA in essay test answers as a result of instruction.

The differential use of topic proportions in the instructional setting has been shown to identify the quality or track the improvement in students' writing (Choi et al., 2017). Choi et al. reported that some students switched in the proportions of topic use from everyday language to academic language following instruction, although the rubric-based scores of their answers did not show any change. Topic proportions, however, also can be used to indicate topics that are highly related to scores. Kwak et al. (2017) found that one of the latent topics, characterized as *discipline specific words*, had a significant positive correlation with the score on the essay test ($r = .70$).

The benefit of LDA is that it provides additional and easy-to-understand information about the content of the essay. Results from Kim et al. (2017) showed that the topic structure from LDA models yielded similar results to a qualitative analysis of the same essays. This suggests that these topic models have the potential to provide constructive feedback in real time to students beyond just a score.

Additional Topic Models

While LSA and LDA are the most commonly used topic models, other topic models have shown promise in modeling text. These are supervised LDA, the bigram model, autoencoders, and word embeddings.

Supervised LDA

Supervised LDA (sLDA; McAuliffe & Blei, 2007) is a variant of LDA that uses a dependent variable to help infer the topics, topic proportions, and topic assignments of the model. The dependent variable is modeled jointly with the LDA model by regressing the dependent variable on the topic proportions (Perrote et al., 2011). The joint modeling framework makes the topic proportions more predictive of the dependent variable. For example, McAuliffe and Blei (2007) found that when trying to predict movie ratings, the sLDA model outperformed the LDA model, followed by regression. Additional covariates, such as document length, can be added to the sLDA, which is referred to as an sLDA with covariates model (Wilcox et al., 2020). When sLDA and the sLDA with covariates model are applied to constructed-response items, the dependent variable is typically a score category or classification of the test taker.

Bigram Topic Model

The bigram topic model (Wallach, 2006) is an extension of the LDA method that incorporates partial order of the words by considering adjacent words. Under the bigram model, each topic is represented as a set of distributions rather than a set of probabilities associated with words. The distribution the topics represent is the combinations of words that occur adjacent to each other. For example, in the sentence *The green ball flew across the sky*, the word combinations of *green ball*, *ball flew*, *flew across*, and so on are considered the bigram terms (i.e., tokens). The bigram topic model uses a similar structure as LDA; however, the model assumes that words will occur adjacently multiple times. That is, essential combinations of words will appear across the multiple documents. For example, it is safe to assume the word *united* and the word *states* will appear next to each other in a single bigram *united states*. If the documents being analyzed do contain words that will appear adjacently across multiple documents, then the bigram topic model will not produce useful information beyond that of the unigram LDA about the text data.

Autoencoder

An autoencoder is a type of artificial neural network used to approximate data by reducing the underlying dimensionality of the data set. The autoencoder extracts features from the data through an unsupervised learning algorithm that contains three major parts: an input layer, a hidden layer, and an output layer. The input layer is a vector representation of the original text, the hidden layer is a reduced vector space, and the output layer is an approximated vector representation of the original text. The length of the vectors from the input and output layers are the same; however, the vectors obtained from the output layer approximate the vectors used in the input layer. There are a variety of autoencoder algorithms for analyzing and extracting features from textual data, including the General Language Understanding Evaluation (A. Wang et al., 2018), the bidirectional encoder representations from transformers (Devlin et al., 2018), and many more (e.g., Y. Chen & Zaki, 2017; Vaswani et al., 2017; Wu et al., 2019; Zhai & Zhang, 2016).

Word2vec and Doc2vec

The word2vec (Mikolov et al., 2013) and doc2vec (Le & Mikolov, 2014) are NLP topic models used in linguistics to find the meaning of words. The word2vec uses one of two algorithms: a continuous “bag-of-words” algorithm and a skip-gram algorithm. These two algorithms are simple neural networks and derive the meaning of words based on the context words surrounding them. The word2vec model assumes that the meaning of words is determined by the words that appear near them. The doc2vec extracts features (i.e., topics) from documents by using the words that appear in them and the meaning of words derived from the word2vec algorithm. In this way, the doc2vec represents an abstract meaning of a document based on the word meanings that word2vec numerically represents. These two models differ from autoencoders in that the order of the text is not considered.

ASSESSING THE VALIDITY OF SCORES THROUGH AUTOMATED ASSESSMENTS

One of the most common uses of topic models is through automated scoring assessments. This section considers how scores predicted by AI assessments may differ from those that humans assign through either a grading or a rating process. Note that human score assignment may not necessarily be the standard through which an AI model is constructed, but for most performance assessments, it is the standard approach. Additionally, it is important to note that general information about validity is addressed elsewhere by Lane and Marion (this volume), and issues related to the use of technology-based assessment are described by Bennett et al. (this volume).

Technology-Based Assessments

Tentative guidelines for technology-based assessments have been drafted by the International Test Commission and Association of Test Publishers (2022), the *Guidelines for Technology-Based Assessments*. This document subsumes what used to be referred to as computer-based testing but now includes testing that might be administered through the Internet, on tablets, on phones, and through speech recognition and the like. These guidelines are based on the *Standards* (AERA et al., 2014) but address issues that may be unique to technology-administered assessments. So, for example, some test takers may lack the background experience to effectively enter their responses on a computer keyboard (a construct-irrelevant factor). The validity concerns in this document are focused on construct underrepresentation and construct-irrelevant factors. The guidelines also cover recommendations for test development, test design and assembly, test delivery environments, scoring, digitally based results reporting, data management, psychometric and technical quality, test security, data privacy, fairness, and accessibility.

Automated Scoring

In his defense of AES, Page (2003) offered the Turing test as a benchmark by which machine-predicted scores might be judged. The test goes something like this: If you hand essays to a scorer behind a screen and cannot tell the difference between those scores assigned by the human scorer and those predicted by a scoring algorithm, then you should treat the two sets of scores as the same. AI-assigned scores generally meet the Turing test (basically a *reliability* argument) but not the test of *transparency*, which most argue is an essential element to establishing the validity of a score (Bejar et al., 2017; Ramineni & Williamson, 2013). The argument here is that if the scoring algorithm gets to the same score using a process different from that of the human scorer, the score will lack some dimension of validity.

The problem is that human scorers can also arrive at the same score using different processes, so the human scoring enterprise is often guided by *rubrics* and *expertise* of the scorer to help define what humans should be attending to when assigning scores. The purpose of the rubric is to define the domain being assessed and articulate the elements

of performance that make it better or worse, complete or incomplete. The challenge is that by restricting the focus of attention to specific aspects of performance, other aspects of performance may be ignored or undervalued (construct underrepresentation). If reliance on expertise is involved, scoring a performance may be influenced by the amount and types of expertise possessed by the human scorer (Wolfe & McVay, 2012). This simply means that AI models based on human scores will inherently possess varying degrees of unreliability. To establish the validity of AI-predicted scores, the twofold goal is to maximize the reliability of scores and conformance to the underlying rubrics or processes used by experts.

Williamson et al. (2012) discussed the interpretational context of using automated scoring in a high-stakes testing environment. The article addressed five areas that should be systematically evaluated before concluding that the use of the technology might be appropriate. These include construct relevance and representation association with human scores, association with independent measures, generalizability of scores, impact on score use, and decisions and consequences. For instance, human raters and machine scoring algorithms may have high agreement on assigned and predicted scores but come to this outcome using different mechanisms. Suppose the machine scoring algorithm is a reliable proxy for the human process under a known range of conditions and meets the criteria listed previously. In that case, it may be acceptable to employ the technology in a high-stakes environment. If it is unreliable and its restrictions cannot be mapped, then the utility of machine scoring will be limited. Given such outcomes, it could even be argued that machine scoring works well for evaluating certain kinds of expository writing for a heterogeneous mix of U.S. school children in Grades 6–12 but does not seem to perform adequately in evaluating poetry or other kinds of stylized writing or does not work well in grades where keyboarding skills have not yet been mastered.

With regard to the evaluation of writing, the current algorithms are used to predict score assignments based on human raters' scores over a variety of academic writing genres. To make significant improvements in the technology, the writing community will have to develop an accepted operational definition of what constitutes "good writing," and computational linguists will have to articulate what they can emulate well and what aspects of writing evaluation require more development.

One common approach to establishing score validity is to employ evidence-centered design (Mislevy et al., 2006). Evidence-centered design is based on a tripartite model that is composed of a construct model, a task, and evidence that "defines rules and criteria for interpreting observations, test taker responses, and processes that can be used to relate observations to the latent constructs in the constructed model and accumulate evidence from responses to tasks" (Ercikan & McCaffrey, 2022). In the case of writing, an argument might be made based on an interpretation of a rubric and the accumulation of sufficient evidence that a score has some particular meaning. The problem is that there are varying degrees of clarity regarding rubrics, processes, and decisions based on the score. (See also ETS, 2021, for a discussion of the collection of validity evidence for the use of automated scores.)

When Rubrics, Processes, and Decisions Are Clear

One of the more popular automated assessment devices used in golf is a technology called Trackman (n.d.). It is usually combined with cameras to show the effectiveness of a golf swing when using specific clubs (e.g., a driver). The Trackman device is placed behind the golfer. It employs Doppler radar to measure multiple aspects of swing performance, including club speed, ball speed, ball spin rate, launch angle, attack angle, swing direction, club path, face angle, ball distance, and ball carry, among others. These attributes are very clearly defined. While there is no perfect golf swing, there are swing models that experts consider as being better than others. For example, if you want to hit a golf ball straighter, you want your swing to align with an ideal swing plane and hit the ball squarely at impact. The goal is not to hit the same shot on every swing, but rather to control the swing to adapt to the situation. For instance, there are times when a golfer might want to shape the ball to go to the right or the left.

The validity issue is whether the measurements taken by Trackman (n.d.) constitute the relevant components of a golf swing and whether the measures taken are accurate and reliable. The accuracy and reliability of the measurements can be confirmed using other tools besides Doppler radar. The relevancy of the measurements can be demonstrated experimentally using golf swing machines. For example, if one wants to hit the ball farther, one can increase the clubhead speed while keeping other “variables” constant. Johansson et al. (2015) used Trackman data with random forest models to accurately predict the handicap (skill level) of a sample of 275 golfers. As is always the case, getting a good golf result with a bad swing is possible, but not likely. It is also possible that there may be other attributes of performance that have not yet been envisioned. Still, for the moment, there seems to be general satisfaction that Trackman and similar instruments result in valid swing assessment scores.

For kinesthetic assessment, there are automated scoring models used in dance (Faridee et al., 2019; Joshi & Chakrabarty, 2021), boxing (Perkins et al., 2017), gymnastics (Shimamoto et al., 2013), and diving (Kramer, 2017), to cite a few areas. These performance domains are similar in that the rubrics for scoring are clear, are well defined, and require expertise or technology to score in real time. The instruments used to aid in the assessment are reliable, but not always perfect. For example, to use automated scoring in boxing, participants must wear specialized jerseys embedded with sensors. During the development of the technology, it was discovered that loose-fitting jerseys resulted in false positives (recording a “hit” when no hit was observed). The experimenters found that in the looser jerseys, the sensors were hitting against each other. The modification ensured that the jerseys had to have a tighter fit (Perkins et al., 2017).

When Rubrics and Processes Are Clear (But Maybe Incomplete) and Decisions Are Muddy

In their treatment of fairness as a component of score validity, Huggins-Manley et al. (2022) illustrated the use of automated video interviews (Hickman et al., 2021) as a mechanism for hiring into a potential job opportunity. During the automated video

interview, the job candidate is presented with a video-taped interviewer, provided with a question or prompt, and then asked to respond. The automated scoring algorithms consider the responses, the length of the answers, the manner in which the response was given, and several other attributes of interaction. Different kinds of employment data were recorded, such as age, race, gender, and level of education, to control for background differences. In the article, Huggins-Manley et al. (2022) advocated using validity arguments to evaluate fairness akin to the kinds of performance statements generated with evidence-centered design. For example, suppose the job in question was someone working as a manager in a call center. In that case, there is no theoretical reason why hiring rates for males and females should be different in terms of their anticipated performance on the automated video interview, assuming that the questions themselves were unbiased. This expectation would result in a formal statement as such.

What Hickman et al. (2021) found, however, was that these expectations can be influenced by confounding effects such as self-selection bias (i.e., the proportion of women and/or men who apply for the job and how to treat those who identify as nonbinary. Surprisingly, Hickman et al. (2021) also found that the scoring technology predicted better decisions when trained interviewers answered as proxies for the applicants rather than the applicants themselves.

When Rubrics, Processes, and Decisions Are Muddy

Shermis (2022) discussed the theoretical misalignment between the kinds of writing assessments that are common in high-stakes tests and those advocated by the writing community. Most accountability writing assessments are based on the composition-rhetoric theory of writing (sometimes referred to as the current-traditional theory of writing; R. E. Young, 1982). The key features of this writing theory include formal correctness, the elegance of style, and the modes of discourse. These are usually implemented as short writing samples suggested by a question, a series of questions, or some other prompt. The result is a writing sample that ranges from 50 to 500 words that may be scored by human raters, machine scoring algorithms, or a combination of the two.

While the composition-rhetoric theory has formed the basis for most accountability exams, the writing community has moved on to other theories of writing in which a specific product is only part of the performance that might be evaluated. For example, the cognitive process theory of writing is aimed at improving and documenting the decisions that writers make (Flower & Hayes, 1981). Participating in a brainstorming exercise or character sketching would be writing activities consistent with a cognitive process theory of writing but would not be directly reflected in a prompt response.

Alternatively, the sociocultural theory of writing stipulates that

writing is a complex social participatory performance in which the writer asserts meaning, goals, actions, affiliations, and identities within a constantly changing, contingently organized social world, relying on shared texts and knowledge. The projection of meaning and shared orientations at a distance requires making assumptions and predictions about who will be reading the texts, what

their interest and knowledge are likely to be, and how they may be using the information. (Bazerman, 2015, p. 18; see also Prior, 2006)

Implicit here is that school, along with many other institutions, is one way a person is socialized, and it is important for the writer to be cognizant of this process. With this theory, writing improves as a function of the interactions (i.e., feedback) the writer receives to achieve a set of norms. Evidence would include measures of improvement (or lack thereof) based on feedback.

The point is that what constitutes writing is different for those in the high-stakes assessment enterprise and those who actually teach writing. The one-shot 40-minute response to a prompt would not capture the evidence that those in the writing community are looking for when they assess the development of someone’s writing performance. Whether the writing is scored by humans or through AI ends up being a secondary issue.

The Transparency of Processes

The major obstacle in establishing the validity of machine score predictions most likely centers on the lack of transparency of processes. This is also a challenge for human raters even when assigning scores based on a rubric. Figure 8.12 shows the process used

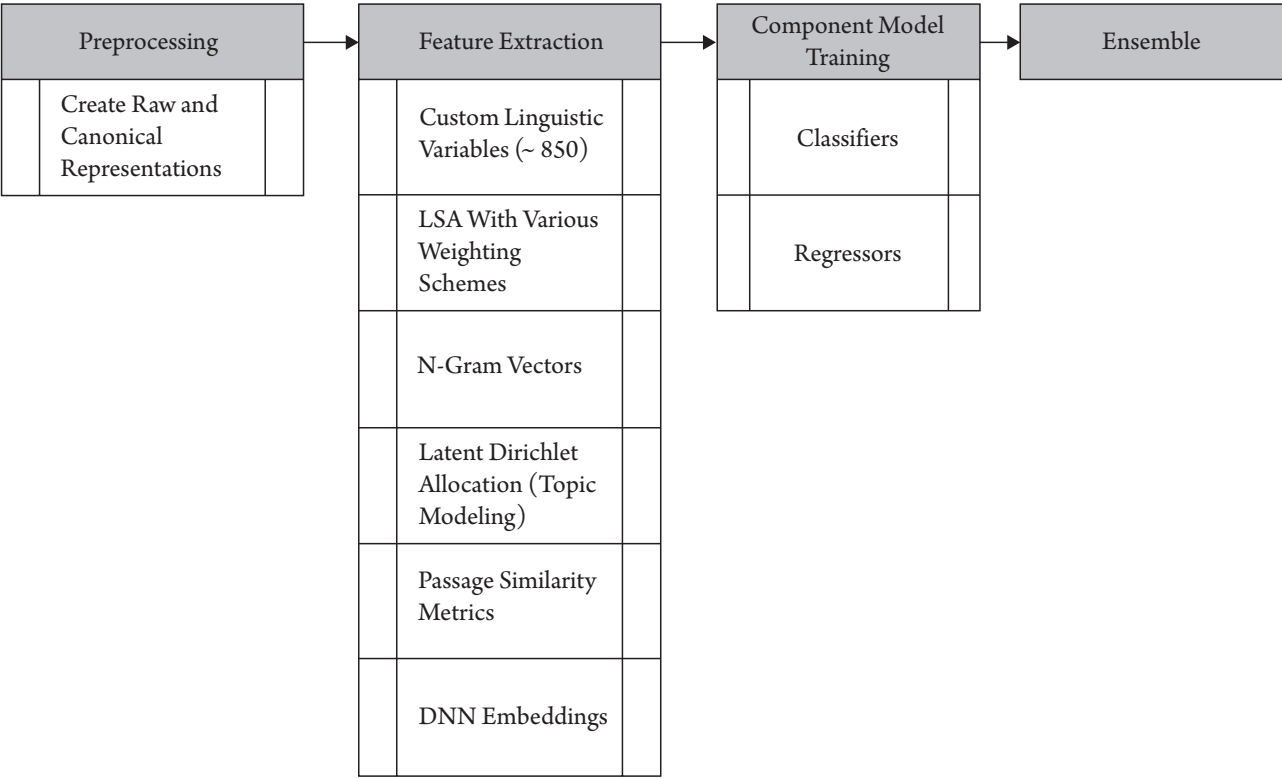


FIGURE 8.12
Project Essay Grade (PEG) Model Training Pipeline

by Measurement, Inc.'s Project Essay Grade (PEG) automated scoring engine. The PEG engine was the 2022 winner of the NAEP contest to score responses to reading items in fourth and eighth grade for the 2017 NAEP test administration (Institute for Education Sciences [National Center for Education Statistics], 2022). In this particular challenge, human raters had $\kappa_w = .91$ averaged across responses to 20 reading items, and PEG score predictions came in at $\kappa_w = .89$. This scoring performance represented a considerable improvement from a similar contest conducted in 2012 (i.e., the ASAP trials).

In 2012, most competitors, including PEG, used parser-tagged linguistic variables, LSA, and the ability to assess passage similarity in what was referred to as a bag-of-words approach (Deerwester et al., 1990) to modeling. Since then, many prediction models have added n-gram vectors, LDA for topic modeling, BERT, and deep neural networks (i.e., machine learning). For example, using the bidirectional encoder representations from transformers allows the language model to be automatically encoded from left to right or right to left. A language model is a type of AI model that is trained to predict the likelihood of the next word in a sequence, given the preceding words in that sequence. It is designed to understand and generate human language and can be used for applications such as speech recognition, machine translation, and text generation. Recent research has shown that a language model that is bidirectionally trained can have a deeper sense of language context and flow than single-direction language or no-direction models (Devlin et al., 2018).

However, outside the field of applied linguistics, these techniques might seem foreign and even unrelated to the process a human rater would go through to assign a score to a writing sample. Shermis (2022) proposed to use either the 6 + 1 Traits or the CEFR (Shermis, 2018) as a way to establish a crosswalk between the concepts employed in AI scoring and the vocabulary used by the writing community to describe their processes. PEG processes align reasonably well with five of the six 6 + 1 Traits (development of ideas, organization, word choice, sentence fluency, conventions). For example, PEG LDA is reasonably well aligned with the trait development of ideas. Without some sort of crosswalk, the measurement and writing communities will find it hard to establish common ground in their description of writing processes.

Finally, Ferrara and Qunbar (2022) suggested that some AI processes will likely remain *unknowable* until we have more experience with machine score prediction algorithms. In refining a score prediction, AI might employ one of many machine learning techniques. Why one of these techniques would perform better on a particular writing sample can be documented but probably not well explained. Much more research is needed to establish the efficacy of some of the tools used in AI score prediction. Still, many parts of the algorithms align well with vetted rubrics used by human rating enterprises. We document elsewhere in this chapter the kinds of possible bias and errors committed by human raters.

Validity and Reliability Evidence for SpeechRater

SpeechRater's score reliability is assessed by first training a multiple linear regression scoring model on a set of about 1,200 responses and corresponding scores assigned by human + expert raters and then evaluating the trained model on an unseen test set of about 300 test takers who have a complete set of six scores (about 1,800 responses total). For each speaker, the sets of six human and automated scores are each summed to obtain an aggregated speaking score. The aggregate speaking scores from human raters and SpeechRater are compared using the Pearson r correlation and quadratic weighted kappa metrics. The correlation between aggregated speaker-level SpeechRater scores and human scores was $r = .61$, and the quadratic weighted kappa was .53. In comparison, the human interrater correlation on the same data set was $r = .63$, and the quadratic weighted kappa was 0.57 (Shermis, Burstein, et al., 2015).

As for possible construct (under-)representation, a validity concern (AERA et al., 2014, chap. 1), the features of SpeechRater are mainly focusing on the construct dimension of "delivery" (e.g., fluency, pronunciation, prosody) and, to a lesser extent, on aspects of "language use" (e.g., vocabulary diversity, grammatical accuracy). Features related to the response content are currently not used in SpeechRater operationally, although recent research has resulted in features associated with this dimension as well (M. Chen & Zechner, 2012; Xie et al., 2012). To represent a comprehensive subset of the construct of speaking proficiency as defined in the TOEFL rubric, a significant amount of additional research and development is still required.

In addition to these considerations of validity, a study on criterion-related validity was conducted by Bridgeman et al. (2012). In this study, 555 U.S. undergraduate students who were native speakers of English were asked to listen to TOEFL iBT responses given by 184 test takers taking one of two distinct forms. The researchers then examined correlations between multiple-choice comprehension scores (based on these listening exercises) and human and SpeechRater scores of the same test takers' responses. They found that the correlation between human expert raters and undergraduate listeners' comprehension scores was higher (.97 and .81 for the two forms, respectively) than for SpeechRater (.55 and .37). The main reason for this discrepancy is very likely that SpeechRater is covering only a subset of the construct, with most of its features focusing on the delivery dimension, few on language use, and none as of yet on topic development.

EMPIRICAL WORK: THE HEWLETT TRIALS

The Hewlett trials, sponsored by the Hewlett Foundation, were conducted in the spring of 2012 as a test of machine scoring for the PARCC and Smarter Balanced consortia (Shermis & Hamner, 2012, 2013). These two entities had responsibility for helping states implement and assess reforms consistent with the Common Core State Standards. The trials consisted of three studies—a commercial vendor demonstration for AES, a worldwide competition for AES, and a worldwide competition for short-form constructed responses.

The first two studies employed eight different essay sets drawn from six states representing the PARCC and Smarter Balanced consortia. In the first two studies, four of the essays were “source based.” A student was asked to read an artifact and then write an essay. The remaining four essays were based on more traditional prompts (i.e., those requiring narrative, descriptive, and persuasive writing). For the first study, over 17,000 essays were randomly divided into two sets, a training set ($n = 13,336$) on which the vendors had 1 month to model the data and a test set for which they were required to make score predictions. The training set essays were each initially scored by two human raters. Significant discrepancies in these scores were then resolved by calling in additional ratings, a so-called resolved score, and the text of the essay. The resolved score reflected the final score assigned by the state. To resolve differences between the two human raters, states applied a set of adjudication rules (e.g., ask a third rater to evaluate the essay) that may or may not have reflected the judgments of the original human raters. The test set consisted only of the text of the essay. Six of the eight essays were transcribed from handwritten documents using one of two transcription services. Accuracy rates were estimated to be over 98%. The challenge to the nine teams was to produce automated scores that best predicted those assigned by human raters.

Performance of automated scoring systems was evaluated on five different measures using the Kaggle platform (onward Kaggle Inc., 2010). The first two measures reflected the distributional properties of the essay scores, mean and standard deviation. The remaining five measures addressed how well the machine scores agreed with those assigned by the human raters (e.g., the resolved score). These included exact agreement, exact + adjacent agreement, kappa, quadratic weighted kappa, and the Pearson product-moment correlation. The AES engines performed well on the distributional measures. With a high degree of consistency, all nine demonstrators in the first study were able to replicate the means and standard deviations for the scores assigned by the human raters. With regard to agreement measures, there was some variability, but the AES engines performed well on three of the five measures (exact + adjacent agreement, quadratic weighted kappa, correlation). On the two measures where the performance was not as high (exact agreement and kappa), there was also high variability among the human raters. In addition, scaling artifacts attributable to the way the state scores were adjudicated may have contributed to the relative lack of precision on predicted scores. In sum, on five of the seven measures, machine scores reasonably approximated those of human raters and, on a few of the data sets, even performed better than their human counterparts. The conclusion of that study was that with additional work, AES could be a viable solution in some aspects of evaluating high-stakes long-form writing assessments. The study did not evaluate the potential for score bias in group differences (i.e., no investigation of differential item functioning), used existing paper-based essays from states as the source of information rather than obtaining text as a computer-administered assessment, and contained a restricted range of evaluation criteria as part of its design.

The second study was very similar to the first, but with some important differences. First, it was conceived as a \$100,000 competition involving 158 teams of data scientists

from around the world. Some of the competitors included a few vendors from the first study who were ineligible to win the monetary prize. Second, the data scientists had 3 months to create and refine their models, whereas the vendors only had 2 weeks. Third, the data sets from the first study were anonymized using the Stanford Named Entity Recognizer (Finkel et al., 2005) so that the likelihood of identifying an individual through persons, organizations, or places was minimized. A study was conducted to determine what, if any, impact might occur through anonymization. On the three measures that were chosen to evaluate impact (mean, standard deviation, quadratic weighted kappa), only minor differences were observed for both NLP- and non-NLP-based scoring engines (Shermis, Lottridge, & Mayfield, 2015). Finally, with the exception of commercial vendors, the competitors were required to produce their algorithms and a white paper that explained how they modeled the data. Results showed that the top international competitors performed slightly better than the commercial vendors. Analysis of the algorithms revealed that the data scientists did not introduce any new techniques, but rather resequenced existing techniques to extract additional explained variance.

The third study compared short-form constructed responses evaluated by both human raters and machine scoring algorithms. As before, both public competitors and commercial vendors vied to develop machine scoring algorithms that would match or exceed the performance of operational human raters in a summative high-stakes testing environment. Data ($N = 25,683$) were drawn from three different states, employed 10 different prompts, and were drawn from two different secondary grade levels. Samples ranging in size from 2,130 to 2,999 were randomly selected from the data sets provided by the states and then randomly divided into three sets: a training set, a test set, and a validation set. Machine performance on all of the agreement measures failed to match that of the human raters. While machine-human agreements were as high as they were for the essay scoring competitions, agreement among human raters was even higher, so that machine performance did not reach the threshold of human raters. The study concluded that additional development had to take place before machine scoring of short-form constructed responses could be commercially viable.

INTERNATIONAL WORK IN MACHINE SCORING

As mentioned at the beginning of this section, there has been an expanded interest in machine scoring beyond the English language. In assessing language differences, three salient dimensions may impact how effectively writing in different languages can be evaluated with automated systems. The first dimension pertains to differences between written and spoken language (e.g., writing complexity), the second dimension addresses dialectical or regional differences within a language (e.g., different words for same object or action), and the third dimension focuses on distinctions among different languages (e.g., different alphabets). Moreover, a few languages are managed by academies (i.e., a formal body admits new words, adjusts the meaning of existing words, and affirms grammatical expressions into the language) while most are not. All of these can impact the efficacy of machine scoring models (Shermis, 2020).

Early efforts focused on using established U.S. English-based scoring engines to score different languages by simply replacing the English parser and tagger with that of the targeted language (Ishioka & Kameda, 2006; Schultz, 2013; Vantage Learning, 2001). Many of the foreign-language scoring engines are, at this point, simply proofs of concept and are typically a combination of a statistical models used in conjunction with LSA or some variant of it (e.g., LDA). They usually employ a weighted linear combination of a subset of parser variables (i.e., metavariables) in addition to a distance metric derived from the LSA component. These tend to work surprising well (i.e., with quadratic weighted kappas sometimes in the low 80s) given that the training set was often small or employed corpora that were only tangentially related to the essay topic. The foreign-language scoring engines are generally an assembly of tools rather than a complete system. As such, they typically require manual handling from one step of the process to the next, rather than as an embedded set of linked steps. Only a few systems included NLP programming to evaluate the more mechanical aspects of writing. Table 8.2 lists several foreign-language AESs.

Table 8.2 Comparison of Non-English-Based Automated Essay Scoring (AES) Systems

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training	Language
Japanese Essay Scoring System	National Center for University Entrance Exams (Japan; Ishioka & Kameda, 2004)	NLP, LSI	Style and content	N/A	External database (newspaper)	Japanese
NiteRater	National Institute for Testing and Evaluation (Israel; Ben-Simon & Cohen, 2011)	Statistical, NLP	Style	N/A	~300	Hebrew
Unnamed (Arabic)	Department of Computer Sciences, Yarmouk University (Jordan; Al-Shalabi, 2016)	Statistical, cosine similarity	Content	N/A	Model answer	Arabic
Filipino Essay Grader	Department of Computer Science, University of the Philippines Diliman (The Philippines; Ong et al., 2011)	Statistical, concept indexing	Content	N/A	150	Tagalog
Unnamed (Finnish)	Department of Computer Science, University of Joensuu (Finland; Kakkonen et al., 2005)	Statistical, LSA	Content	N/A	143, external database	Finnish
Unnamed (Thai)	(Loraksa & Peachavanish, 2007)	Artificial neural network, LSA	Content	N/A	40	Thai

(continued)

Table 8.2 (continued)

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training	Language
Unnamed (Chinese)	Department of Computer Science and Information Engineering, National Kaohsiung University (Taiwan; Chang & Lee, 2009)	Statistical, R-chain similarity (paragraph)	Content	N/A	574	Chinese
Unnamed (Chinese)	(Yao, 2013)	Latent-semantic vector space models	Content	N/A	970	Chinese
Unnamed (Korean)	Korea Maritime and Ocean University (Korea; Cheon et al., 2015)	Statistical	Content (short answer)	N/A	Model answer	Korean
Norwegian Learner Language Corpus	University of Oslo (Berggren et al., 2019)	Statistical, neural, convolutional, and recurrent networks	Content (essays)	N/A	Unknown	Norwegian
Korean Automatic Scoring System	School of Computer Science, Kookmin University (Korea; Jang et al., 2014)	Statistical	Content (short answer)	N/A	Model answer	Korean
Unnamed (Swedish)	Institute for Linguistics, University of Stockholm (Sweden; Smolentzov, 2013)	Statistical, NLP	Content	N/A	External database, 600 essays	Swedish
Unnamed, a collection of tools (French)	University of Paris (France; Parslow, 2015)	Statistical, NLP	Content	N/A	External database	French
Unnamed, a collection of tools (English and French)	Center for Research in Applied Measurement and Evaluation, University of Alberta (Canada; Latifi et al., 2014)	Statistical	Content	N/A	Not Reported	English French
Automated Essay Grading With Latent Semantic Analysis	Department of Information Systems and New Media, Vienna University (Austria; Hagerman, 2011)	Statistical, LSA	Content	N/A	Model answer	German

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training	Language
Freeling, a collection of tools (Spanish and Basque)	(Lloberes et al., 2010)	Statistical	Style	N/A	30	Spanish Basque
Unnamed (Chinese)	(Xu & Liu, 2016)	Statistical, LSA (Coh-Metrix)	Content and style	N/A	Not reported	Chinese
Reader-Bench (Dutch)	(Dascalu et al., 2017)	Statistical, LSA	Content	NA	External database	Dutch

Note. Adapted from “International Applications of Automated Scoring,” by M. D. Shermis, in D. Yan, A. A. Rupp, and P. W. Foltz (Eds.), *Handbook of Automated Scoring: Theory Into Practice* (Table 7.1), 2020, CRC Press. Reprinted with permission.

FEEDBACK FOR FORMATIVE ASSESSMENT

Ellis Page originally envisioned AES as a “teacher’s helper” to provide the extensive feedback that classroom teachers did not have the time to give when faced with grading 125 papers (5 classes × 25 students/class) for each writing assignment (Page et al., 1994). Receiving ample and significant feedback in time to improve writing performance is closely related to improved levels of writing performance (Biber et al., 2011; Landauer et al., 2009). Recent meta-analyses of high-quality research studies have identified writing instruction practices that can best improve student reading and writing performance. The studies—*Reading Next* (Biancarosa & Snow, 2004), *Writing Next* (S. Graham & Perin, 2007), *Writing to Read* (S. Graham & Hebert, 2010), *Informing Writing* (S. Graham et al., 2011), and *The Effectiveness of Feedback for L1-English and L2-Writing Development* (Biber et al., 2011)—provide a range of practices, as well as the statistical effect size (an estimate of the magnitude of the instruction) on student learning, that may serve as a guide to best instructional practices: providing feedback, developing inquiry activities for a particular writing task, encouraging self-assessment, using student evaluation of writing using scales, fostering sentence combining instruction, monitoring of writing progress by the instructor, study and emulation of good models of writing, and promoting free writing. There are also cautions regarding allowing students to write using paper and pencil when they are inexperienced in word processing and the exclusive use of grammar instruction that has not been demonstrated to be an effective treatment. Such empirical findings have allowed the development of a set of best principles for writing instruction: the creation of a supportive writing environment, instruction in writing strategies, knowledge acquisition of effective writing, and attention to foundational skills (S. Graham et al., 2013).

CONCLUSION

The research and work discussed in this chapter have demonstrated the profound impact that technology has had on test administration and scoring. The technology extends from how we administer and score assessments to the detection of possible cheating. Some of these developments have opened adaptations for individuals with special needs in a way that was not possible with traditional paper-and-pencil tests. However, the scoring technology does not always follow the same process or logic human evaluators use when making judgments and sometimes lacks transparency. It is incumbent on those who develop automated scoring processes and those who use them to ensure that the technology does not introduce construct-irrelevant variance in interpreting test scores. Guidelines have been developed to help those engaged in this enterprise, but it is always incumbent to provide evidence that the resulting test scores are appropriately interpreted.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Susan Lottridge and Peter Foltz for their insightful comments and constructive suggestions. We also appreciate the thoughtful recommendations and remarkable editorial work of the book's coeditors, Drs. Linda Cook and Mary Pitoniak.

REFERENCES

- Abt, C. C. (1970). *Serious games*. Viking Press.
- Alayed, H., Frangoudes, F., & Neuman, C. (2013, August 11–13). Behavioral-based cheating detection in online first person shooters using machine learning techniques. *Proceedings of the 2013 IEEE Conference on Computational Intelligence in Games* (pp. 1–8). <https://doi.org/10.1109/cig.2013.6633617>
- Al-Shalabi, E. F. (2016). An automated system for essay scoring of online exams in Arabic based on stemming techniques and Levenshtein edit operations. *International Journal of Computer Science Issues*, 13(5), 45–50. <https://doi.org/10.20943/01201605.4550>
- Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2), 133–149. <https://doi.org/10.1109/tsmcc.2011.2134847>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573.
- Annetta, L. A. (2008). *Serious educational games: From theory to practice*. Sense Publishers.

- Annetta, L. A., & Bronak, S. C. (2011). *Serious educational games: Practical methods and models for educational games, simulations and virtual worlds*. Sense Publishers.
- Attali, Y. (2019). Rater certification tests: A psychometric approach. *Educational Measurement: Issues and Practice*, 38(2), 6–13.
- Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, 10(3), 1–16. <http://jtla.bc.edu>
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3), 1–29. <http://www.bc.edu/research/intasc/jtla/journal/v4n3.shtml>
- Banks, J. (1998). Principles of simulation. In J. Banks (Ed.), *Handbook of simulation: Principles, methodology, advances, applications and practice* (Vol. 12, pp. 3–30). Wiley-Interscience.
- Bazerman, C. (2015). What do sociocultural studies of writing tell us about learning to write? In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 11–23). Guilford Press.
- Beigman-Klebanov, B., Madnani, N., & Burstein, J. (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics*, 1, 99–110. <http://www.transacl.org/papers/>
- Bejar, I. I., Mislevy, R. J., & Zhang, M. (2017). Automated scoring with validity in mind. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 226–246). John Wiley & Sons.
- Ben-Simon, A., & Cohen, Y. (2011). *The Hebrew Language Project: Automated essay scoring & readability analysis* [Paper presentation]. International Association for Educational Assessment Annual Conference, Manila, Philippines.
- Berggren, S. J., Rama, T., & Øvrelid, L. (2019, Aug 2, 2019). Regression or classification? Automated essay scoring for Norwegian. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 92–102). Florence, Italy: ACL. <https://aclanthology.org/W19-4400>
- Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology*, 63(3), 301–308.
- Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review*, 6(2), 205–212.
- Bhat, S., & Yoon, S.-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, 67, 42–57.
- Biancarosa, G., & Snow, C. E. (2004). *Reading Next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York*. Alliance for Excellent Education.
- Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis* (TOEFL iBT Research Report No. iBT-14). ETS. <https://doi.org/10.1002/j.2333-8504.2011.tb02241.x>

- Blackburn, J., Kourtellis, N., Skvoretz, J., Ripeanu, M., & Iamnitchi, A. (2014). Cheating in online games: A social network perspective. *ACM Transactions on Internet Technology*, 13(3), 1–25. <https://doi.org/10.1145/2602570>
- Blei, D. M., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, 27(6), 55–65. <https://doi.org/10.1109/msp.2010.938079>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education*, 25(3), 141–152. <https://doi.org/10.1177/07419325040250030201>
- Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvements. In E. M. Airoldi, D. M. Blei, E. A. Erosheva, & S. E. Fienberg (Eds.), *Handbook of mixed membership models and their applications* (pp. 225–255). Chapman & Hall/CRC.
- Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, 29(1), 91–108. <https://doi.org/10.1177/0265532211411078>
- Burdick, H., Swartz, C., Stenner, J., Fitzgerald, J., Burdick, D., & Hanlon, S. (2011, November 30–December 3). *A computer-analytic writing ability scale: A validity study* [Paper presentation]. Literacy Research Association Annual Meeting, Jacksonville, FL, United States.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April 14–16). *Computer analysis of essays* [Paper presentation]. National Council of Measurement in Education Annual Meeting, San Diego, CA, United States.
- Burstein, J., Tetreault, J., & Chodorow, M. (2013). Holistic annotation of discourse coherence quality in noisy essay writing. *Dialogue and Discourse*, 4(2), 34–52. <https://doi.org/10.5087/dad.2013.202>
- Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater[®] automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). New York: Routledge.
- Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S., Choi, Y. J., & Cohen, A. S. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of the language of scientific investigation practices. *Double Helix: A Journal of Critical Thinking and Writing*, 2, 1–21.
- Chang, T. H., & Lee, C. H. (2009, Dec 7–9). Automatic Chinese essay scoring using connections between concepts in paragraphs. *Proceedings of the International Conference on Asian Language Processing*, (pp. 265–268), Singapore: IEEE Computer Society. <http://doi.org/10.1109/IALP.2009.63>
- Chemudugunta, C. (2010). *Text mining with probabilistic topic models: Applications in information retrieval and concept modeling*. Lambert Academic Publishing.
- Chen, H. (2021, March 15). *Does word error rate matter?* <https://www.smartaction.ai/blog/does-word-error-rate-matter/>. SmartAction

- Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., & Yoon, S.-Y. (2016, Nov 12-16, 2016). Automated scoring of interview videos using Doc2Vec multimodal feature extraction paradigm. *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 161–168), Tokyo, Japan. <https://doi.org/10.1145/2993148.2993203>
- Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. *Proceedings of the Twelfth Annual Conference of the International Speech Communication Association* (pp. 1861–1864). Florence, Italy: ISTCA. https://www.isca-archive.org/interspeech_2011/
- Chen, M., & Zechner, K. (2012, June 7). Using an ontology for improved automated content scoring of spontaneous non-native speech. *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2012* (pp. 86–94). Montreal, Canada: ACL. <https://aclanthology.org/W12-2010.pdf>
- Chen, Y., & Zaki, M. J. (2017, August 13–17). Kate: K-competitive autoencoder for text. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 85–94). Halifax, NS, Canada: ACM. <https://doi.org/10.1145/3097983.3098017>
- Cheng, J., D'Antilio, Y. Z., Chen, X., & Bernstein, J. (2014, June 26). Automatic assessment of the speech of young English learners. *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 12–21, ACL. <https://doi.org/10.3115/v1/W14-18>
- Cheon, M. A., Seo, H. W., Kim, J. H., Noh, E. H., & Sung, K. H. (2015, July 31). An automated scoring tool for Korean short-answer questions based on semi-supervised learning. *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 59–63). Beijing, China: ACL & AFNLP. <https://doi.org/10.18653/v1/W15-44>
- Choi, H. J., Kwak, M., Kim, S., Cohen, A. S., & Bottge, B. A. (2017, July 18–21). An application of a topic model to two educational assessments. *Proceedings of the annual meeting of the Psychometric Society* (pp. 449–459), Zurich, Switzerland. Springer Verlag.
- Cizek, G. J., & Wollack, J. A. (2017). Exploring cheating on tests. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 3–19). Routledge.
- Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 5(2), 115–124.
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities Research and Practice*, 20(4), 225–233. <https://doi.org/10.1111/j.1540-5826.2005.00138.x>
- Cohen, A. S., & Wollack, J. A. (2006). Test scoring and administration. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). ACE/Praeger.
- College Board. (2021). *AP Music Theory audio prompts and sample responses: 2021*. Retrieved June 6, 2022 from <https://apcentral.collegeboard.org/courses/ap-music-theory/exam/music-theory-audio-prompts-2021>

- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe. (2002). *Common European framework of reference for languages: Learning, teaching, assessment*.
- Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017, September 12–15). Reader Bench learns Dutch: Building a comprehensive automated essay scoring system for Dutch language. *Proceedings of the 12th European Conference on Technology Enhanced Learning, EC-TEL 2017* (pp. 606–609), Tallin, Estonia. Springer.
- Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(sici\)1097-4571\(199009\)41:6](https://doi.org/10.1002/(sici)1097-4571(199009)41:6)
- Dendir, S., & Maxwell, R. S. (2020). Cheating in online courses: Evidence from online proctoring. *Computers in Human Behavior Reports*, 2, 100033.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805
- Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Moere, A. V. (2008). Evaluation of the usefulness of the Versant for English Test: A response. *Language Assessment Quarterly*, 5(2), 160–167. <https://doi.org/10.1080/15434300801934744>
- DuBois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.
- Dumais, S. T. (1994, Nov 2–44). Latent semantic indexing (LSI). *Proceedings of the Third Text Retrieval Conference (TREC-3)* (pp. 218–219), Gaithersburg, MD. National Institute of Standards and Technology.
- Duong, E. V., Mellom, P., & Hixon, R. (2019, March 9–12). *Using topic modeling to analyze the effects of instructional conversation on 3rd grade students' writing* [Paper presentation]. American Association for Applied Linguistics Annual Meeting, Atlanta, GA, United States.
- Ejima, C., & Takeuchi, K. (2022, July 2–8). Statistical learning models for Japanese essay scoring toward one-shot learning. *Proceedings of the 12th International Congress on Advanced Applied Informatics* (pp. 313–318), Kanazawa, Japan. IEEE. <https://doi.org/10.1109/IIAIAAI55812.2022.00070>
- Elliot, S. M. (2003). Intellimetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. 71–86). Lawrence Erlbaum Associates.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93–112. <http://www.jstor.org/stable/pdfplus/1435170.pdf?acceptTC=true>
- Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Lawrence Erlbaum Associates.

- Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3), 272–287. <https://doi.org/10.1111/jedm.12332>
- ETS. (2013). *TOEFL® speaking rubric*. Retrieved April 19, 2013, from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf
- ETS. (2021). *Best practices for constructed-response scoring*. https://www.ets.org/content/dam/ets-org/pdfs/about/cr_best_practices.pdf
- Evanini, K., & Zechner, K. (2020). Overview of automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Routledge.
- Faridee, A. Z. M., Ramamurthy, S. R., & Roy, N. (2019). HappyFeet. *GetMobile: Mobile Computing and Communications*, 22(3), 10–16. <https://doi.org/10.1145/3308755.3308759>
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3), 288–313.
- Finkel, J. R., Grenager, T., & Manning, C. (2005, June 25–30). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, (pp. 363–370). ACL. <https://aclanthology.org/P05-1.pdf>
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387. <https://doi.org/10.2307/356600>
- Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Organizational Process*, 25(2–3), 285–307.
- Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligence essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2), 1–6. <http://imej.wfu.edu/articles/1999/2/04/>
- Fremer, J. J., & Ferrara, S. (2013). Security in large-scale paper and pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 31–52). Routledge.
- Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21, 353–367.
- Galli, L., Loiacono, D., Cardamone, L. I., & Lanzi, P. L. (2011, September 11–14). A cheating detection framework for Unreal Tournament III: A machine learning approach. *Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Games* (pp. 266–262). Seoul, South Korea. IEEE. <https://doi.org/10.1109/CIGI18774.2011>
- Ghai, W., & Singh, N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications*, 41(8), 42–50.
- Gierl, M. J., Lai, H., & Matovinovic, D. (2020). Augmented intelligence and the future of item development. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (Vol. 1, pp. 1–24). Information Age Publishing.

- Glazer, N., & Wolfe, E. W. (2020). Understanding and interpreting human scoring. *Applied Measurement in Education*, 33(3), 191–197.
- Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In F. L. Bauer (Ed.), *Linear algebra: Handbook for automatic computation* (Vol. 2, pp. 134–151). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-39778-7_10
- Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings: A Report from the Center for Education Compensation Reform*. CERC.
- Graham, S., Harris, K., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment. A Report from the Carnegie Corporation of New York*. Carnegie Corporation.
- Graham, S., & Hebert, M. (2010). *Writing to read: Evidence for how writing can improve reading: A report from the Carnegie Corporation of New York*. Carnegie Corporation of New York.
- Graham, S., MacArthur, C. A., & Fitzgerald, J. (2013). *Best practices in writing instruction*. Guilford Press.
- Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools: A Report to the Carnegie Corporation of New York*. Carnegie Corporation.
- Green, A. (2013). *Relating the Versant English Placement Test to the Common European Framework of Reference*. CRELLA.
- Gregg, N., & Nelson, J. M. (2010). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, 45(2), 128–138. <https://doi.org/10.1177/0022219409355484>
- Guo, X., & Lei, P.-W. (2020). Effect of quality characteristics of peer raters on rating errors in peer assessment. *International Journal of Testing*, 20(3), 206–230.
- Hagerman, C. (2011). An evaluation of automated writing assessment. *The Jaltcall Journal*, 7(3), 271–292.
- Hauenstein, N. M., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, 25(3), 253–266.
- Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323–1351. <https://doi.org/10.1037/apl0000695>
- Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145–159. <https://doi.org/10.1017/s1351324906004189>
- Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language*, 25(2), 282–306. <https://doi.org/https://doi.org/10.1016/j.csl.2010.06.001>
- Howson, P. (2013). *The English effect*. British Council.

- Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). *Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training*. <https://doi.org/10.48550/arXiv.2104.01027>
- Huggins-Manley, A. C., Booth, B. M., & K., D. M. S. (2022). Toward argument-based fairness with an application to AI-enhanced educational assessments. *Journal of Educational Measurement*, 59(3), 362–388. <https://doi.org/10.1111/jedm.12334>
- inkle. (2014). *80 days* [Screenshot].
- Institute for Education Sciences [National Center for Education Statistics]. (2022, January 21). *Four teams win top prize in automated scoring challenge for the nation's report card* [Press release].
- International Test Commission & Association of Test Publishers. (2022). *Guidelines for technology-based assessment (draft)*. ITC.
- Ishioka, T., & Kameda, M. (2004, August 30–September 44). Automated Japanese Essay Scoring System: JESS. Proceedings of the 15th International Workshop on Database and Expert Systems Applications (pp. 4–8), Zaragoza, Spain. Springer.
- Ishioka, T., & Kameda, M. (2006, July 17–18). Automated Japanese Essay Scoring System based on articles written by experts. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 233–240), Sydney, Australia. ACL. <https://aclanthology.org/P06-1030.pdf>
- Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology*, 64(5), 502.
- Jang, E.-S., Kang, S.-S., Noh, E.-H., Kim, M.-H., Sung, K.-H., & Seong, T.-J. (2014, April 1–3). KASS: Korean Automatic Scoring System for short-answer questions. *Proceedings of the 6th International Conference on Computer Supported Education* (pp. 226–230), Barcelona, Spain. INSTICC.
- Jensema, C. J. (1974). The validity of Bayesian tailored testing. *Educational and Psychological Measurement*, 34, 757–766.
- Johansson, U., Konig, R., Brattberg, P., Dahlbom, A., & Riveiro, M. (2015, Dec 7–95). Mining TrackMan golf data. *Proceedings of the International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 380–385). IEEE. <https://doi.org/10.1109/CSCI.2015.77>
- Joshi, M., & Chakrabarty, S. (2021). An extensive review of computational dance automation techniques and applications. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2251), 20210071. <https://doi.org/10.1098/rspa.2021.0071>
- Kakkonen, T., Myller, N., & Sutinen, E. (2006, October 16–19). Applying part-of-speech enhanced LSA to automatic essay grading. *Proceedings of the 4th IEEE International Conference on Information Technology: Research and Education* (pp. 110–120), Tel Aviv, Israel. IEEE. <https://doi.org/10.1109/IEEECONF11358.2006>

- Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005, Jun 29, 2005). Automatic essay grading with probabilistic latent semantic analysis. *Proceedings of the Second Workshop on Building Educational Applications Using NLP* (pp. 29-36). ACL.
- Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C., & Cohen, A. S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics*, 1(1), 82–102. <https://doi.org/10.37514/JWA-J.2017.1.1.05>
- Kramer, R. S. S. (2017). Sequential effects in Olympic synchronized diving scores. *Royal Society Open Science*, 4(1), 160812. <https://doi.org/10.1098/rsos.160812>
- Krath, J., Schürmann, L., & von Korflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior*, 125, 106963. <https://doi.org/https://doi.org/10.1016/j.chb.2021.106963>
- Kwak, M., Kim, S., & Cohen, A. S. (2017, January 12–13). *Latent Dirichlet analysis of text constructed response answers* [Paper presentation]. 4th Annual Writing Analytics Conference, St. Petersburg, FL, United States.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284.
- Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 412–417). Lawrence Erlbaum Associates.
- Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice*, 48(1), 44–52. <https://doi.org/10.1080/00405840802577593>
- Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis* (T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch, Eds.). Psychology Press.
- Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming*, 45(6), 752–768.
- Latifi, S., Gierl, M. J., Boulais, A. P., & De Champlain, A. F. (2014). Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. *Evaluation & the Health Professions*, 39(1), 100–113. <https://doi.org/10.1177/0163278715605358>
- Lazarin, M. (2006). *Improving assessment and accountability for English language learners in the No Child Left Behind Act. Report 6*. National Council of La Raza. <http://publications.nclr.org/handle/123456789/1091>
- Le, Q., & Mikolov, T. (2014, June 21–264). *Distributed representations of sentences and documents*. Proceedings of the 31st International Conference on Machine Learning (Vol 32(1), pp. 1188–1196), Beijing, China. PRML Press.
- Leacock, C., & Chodorow, M. (2003). Automated grammatical error detection. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 195–208). Lawrence Erlbaum Associates.

- Lilja, M. (2018). *Automatic essay scoring of Swedish essays using neural networks* [Unpublished dissertation]. Uppsala University. <https://www.diva-portal.org/smash/get/diva2:1213688/FULLTEXT01.pdf>
- Lloberes, M., Castellón, I., & Padró, L. (2010, May 17–23). Spanish Freeing dependency grammar. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) (pp. 693–699), Valletta, Malta. ELRA.
- Loewus, L. (2016, October 10). Why don't the Common-Core Standards include cursive writing? *Education Week*. <https://www.edweek.org/teaching-learning/why-dont-the-common-core-standards-include-cursive-writing/2016/10>
- Loraksa, C., & Peachavanish, R. (2007, March 27–30,). *Automatic Thai-language essay scoring using neural network and latent semantic analysis*. First Asia International Conference on Modelling & Simulation (pp. 400–402), Phyket, Thailand.: IEEE. <https://ieeexplore.ieee.org/document/4148694>
- Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1(1), 95–100.
- Lottridge, S., Burkhardt, A., & Boyer, M. (2020a). Digital module 18: Automated scoring <https://ncme.elevate.commpartners.com>. *Educational Measurement: Issues and Practice*, 39(3), 141–142.
- Lottridge, S., Godek, B., Jafari, A., & Patel, M. (2020b). Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies [White paper]. <https://files.portal.cambiumast.com/corporate-site/documents/CAI-Cambium-Comparing-Robustness-Automated-Scoring-Approaches.pdf>
- Lottridge, S. M., Schultz, E. M., & Mitzel, H. C. (2013). Using automated scoring to monitor reader performance and detect reader drift in essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 233–250). Routledge.
- Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities. *Review of Educational Research*, 80(4), 611–638. <https://doi.org/10.3102/0034654310364063>
- MacArthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of Special Education*, 21(3), 22–42. <https://doi.org/10.1177/002246698702100304>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Mardones-Segovia, C. A., Wheeler, J. M., Choi, H. J., Wang, S., & Cohen, A. S. (2023). Model selection for latent Dirichlet allocation in assessment data. *Psychological Test and Assessment Modeling*, 64(1), 3–35.
- Marini, J. P., Westrick, P. A., Young, L., Ng, H., Shmueli, D., & Shaw, E. J. (2019). *Differential validity and prediction of the SAT®: Examining first-year grades and retention to the second year* College Board.
- Mayfield, E., & Black, A. W. (2020, July 10). Should you fine-tune BERT for automated essay scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for*

- Building Educational Applications* (pp. 151–162), ACL. <https://aclanthology.org/2020.bea-1/>
- Mayfield, E., & Rosé, C. (2013). LightSIDE: Open source machine learning for text. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124–135). Routledge. <https://doi.org/10.4324/9780203122761.ch8>
- Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from conventional and automated educational and psychological tests: A review of literature* (College Board Technical Report No. 88-21). College Board. <https://doi.org/10.1002/j.2330-8516.1988.tb00277.x>
- McAuliffe, J. D., & Blei, D. M. (2007). Supervised topic models. In J. Platt, D. Kohler, Y. Singer, & S. Roweis (Eds.), *Advances in 20* (pp. 121–128). NeurIPS. <http://papers.nips.cc/paper/3328-supervised-topic-models.pdf>
- McClellan, C. A. (2010). Constructed-response scoring—doing it right. *R&D Connections*, 13, 1–7.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511894664.020>
- Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing* (Law School Admission Council Technical Report No. 03-03). Law School Admission Council.
- Microsoft Corporation. (2022). *Computer Vision v4.0*. extension://bfgdogplmn didlpjfhiojckpakkdjkkil/pdf/viewer.html?file=https%3A%2F%2Flearn.microsoft.com%2Fpdf%3Furl%3Dhttps%253A%252F%252Flearn.microsoft.com%252Fen-us%252Fazure%252Fcognitive-services%252Fcomputer-vision%252Ftoc.json
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013, December 5–8). Distributed representations of words and phrases and their compositionality. *Proceedings of the Advances in Neural Information Processing Systems* (Vol 26, pp. 1188–1196) NeurIPS.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2005). *Computer-based testing: Building the foundation for future assessments*. Routledge.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–47). Routledge. <https://doi.org/10.4324/9780415963572>
- Moon, T. R., & Hughes, K. R. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice*, 21(2), 15–19.
- Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspectives. *Journal of Educational Computing Research*, 43(4), 511–531. <https://doi.org/10.2190/ec.43.4.e>
- Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4), 371–389.

- Noonan, L. E., & Sulsky, L. M. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, 14(1), 3–26.
- Ong, D. A. N., Razon, A. R., Perigrino, J. M., & Guevara, R. C. L. (2011, March 11). Automated Filipino Essay Grader with concept-indexing [Paper presentation]. Presented at the 11th Philippine Computing Science Conference, Naga City, Philippines.
- onward Kaggle Inc. (2010). *Home page*. Retrieved July 18, 2014 from <http://www.kaggle.com>
- Ormerod, C., Lottridge, S., Harris, A. E., Patel, M., van Wamelen, P., Kodeswaran, B., Woolf, S., & Young, M. (2022). Automated short answer scoring using an ensemble of neural networks and latent semantic analysis classifiers. *International Journal of Artificial Intelligence in Education*, 3, 1–30. <https://doi.org/10.1007/s40593-022-00294-2>
- Östling, R., Smolentzov, A., Hinnerich, B., Högl, E. (2013, June 13). *Automated essay scoring for Swedish*. Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 42–47), ACL.
- Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates.
- Page, E. B., Truman, D. L., & Lavoie, M. J. (1994). *“Teacher’s Helper”: Proposed use of Project Essay Grade for the English classroom*. South Atlantic Modern Language Association.
- Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Springer. <https://doi.org/10.1007/0-306-47531-6>
- Parshall, C. G., & Harnes, J. C. (2014). Improving the quality of innovative item types: Four tasks for design and development. *Journal of Applied Testing Technology*, 10(1), 1–20.
- Parslow, N. L. (2015). *Automated analysis of L2 French writing: A preliminary study* [Unpublished masters thesis]. University of Paris Diderot. https://www.researchgate.net/profile/Nicholas-Parslow/publication/279715750_Automated_Analysis_of_L2_French_Writing_a_preliminary_study/links/559834a008ae5d8f3933efd6/Automated-Analysis-of-L2-French-Writing-a-preliminary-study.pdf
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341–384.
- Pearson. (2017). *AZELLA Arizona English Language Learner Assessment*. Arizona Department of Education.
- Perkins, P., Helmer, R. J. N., Mackintosh, C., & Hahn, A. G. (2017). Reflections on long-term development and use of automated scoring technology in a sport (modified

- boxing) context. *World Journal of Engineering and Technology*, 5(3), 455–480. <https://doi.org/10.4236/wjet.2017.53040>
- Perrote, A. J., Wood, F., Elhadad, N., & Bartlett, N. (2011). Hierarchically supervised latent Dirichlet allocation. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 24* (pp. 2609–2617). NeurIPS.
- Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher*, 40(3), 103–116.
- Prior, P. (2006). A sociocultural theory of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 54–66). The Guilford Press.
- Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research*, 58(4), 771–790. <https://doi.org/10.1177/0735633119881472>
- Ramanarayanan, V., Leong, C. W., Chen, L., Feng, G., & Suendermann-Oeft, D. (2015, November 9–13). Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. *Proceedings of the 2015 Association for Computing Machinery on International Conference on Multimodal Interaction* (pp. 25–30), ACM.
- Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25–39. <https://doi.org/10.1016/j.asw.2012.10.004>
- Ricker-Pedley, K. L., Hines, S., & Connelly, C. (2020). Operational human scoring at scale. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds), *Handbook of automated scoring* (pp. 171–194). Chapman & Hall/CRC. <https://doi.org/10.1201/9781351264808>
- Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503–520. <https://doi.org/10.1108/00220410410560582>
- Roch, S. G., & O'Sullivan, B. J. (2003). Frame of reference rater training issues: Recall, time and behavior observation training. *International Journal of Training and Development*, 7(2), 93–107.
- Rosyid, H. A. (2018). *Adaptive serious educational games using machine learning*. University of Manchester.
- Saka, N., Kleper, D., & Kennet-Cohen, T. (2022). Assessing fairness in selection toward applicants who request accommodations in higher education admissions tests. *Assessment in Education: Principles, Policy & Practice*, 29(4), 441–461.
- Schmidt, F. L., Urry, V. W., & Gugel, J. F. (1978). Computer assisted tailored testing: Examinee reactions and evaluations. *Educational and Psychological Measurement*, 38(2), 265–273. <https://doi.org/10.1177/001316447803800208>
- Schofield, A., Magnusson, A., & Mimno, D. (2017, April 3–7). Understanding text pre-processing for latent Dirichlet allocation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (Vol 2, pp. 432–436), Valencia, Spain. Vancouver, BC, Canada: ACL.

- Schultz, M. T. (2013). The IntelliMetric™ automated essay scoring engine—a review and an application to Chinese essay scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 89–98). Routledge.
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>
- Shermis, M. D. (2018). Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring. *Applied Measurement in Education*, 31(3), 177–190.
- Shermis, M. D. (2020). International applications of automated essay scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 113–132). CRC Press.
- Shermis, M. D. (2022). Anchoring validity evidence for automated essay scoring. *Journal of Educational Measurement*, 59(3), 314–337. <https://doi.org/10.1111/jedm.12336>
- Shermis, M. D. (2024). Automated scoring for NAEP short-form constructed responses in reading. In M. D. Shermis & J. Wilson (Eds.), *Routledge international handbook of automated essay evaluation* (pp. 117–140). Routledge.
- Shermis, M. D., Burstein, J., Brew, C., Higgins, D., & Zechner, K. (2015). Recent innovations in machine scoring of student and test taker written and spoken responses. In S. L. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed.; pp. 335–354). Routledge. <https://doi.org/10.1002/ets2.12096/full>
- Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 1–15). Routledge.
- Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art in the machine scoring: Analysis. National Council on Measurement in Education.
- Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 298–312). Routledge.
- Shermis, M. D., Lottridge, S., & Mayfield, E. (2015). The impact of anonymization for automated essay scoring. *Journal of Educational Measurement*, 52(4), 419–436.
- Shimamoto, H., Taki, T., & Hasegawa, J. (2013). A Study of an automated scoring system for the twist skill in horizontal bar of artistic gymnastics. *ECTI Transactions on Computer and Information Technology*, 7(1), 1–6.
- Shin, J., Guo, Q., & Gierl, M. J. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology*, 10, 825.
- Shivakumar, N., & Garcia-Molina, H. (1995). The SCAM approach to copy detection in digital libraries. *D-Lib Magazine*, 1(5), 11–13. <http://doi.org/10.1045/november95-shivakumar>

- Simon, J. P. (2019). Artificial intelligence: Scope, players, markets and geography. *Digital Policy, Regulation and Governance*, 21(3), 208–237.
- Sireci, S. G., Li, S., & Scarpatti, S. (2003). *The effects of test accommodations on test performance: A review of the literature* (CEAR Report No. 485). University of Massachusetts Amherst.
- Sireci, S. G., Scarpatti, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490.
- Smolentzov, A. (2013). *Automated essay scoring: Scoring essays in Swedish*. Institute of Linguistics Report. University of Stockholm. <https://www.diva-portal.org/smash/get/diva2:602025/FULLTEXT01.pdf>
- Spencer, J. (2019, October 14). *Writing prompt: You have just been shrunk down to two inches tall* [Video] YouTube. <https://www.youtube.com/watch?v=Dzhxq3nCCqU>
- Steam. *Dissection Simulator* [Screenshot]. 2023, April 11. https://store.steampowered.com/app/1046910/Dissection_Simulator_Frog_Edition/
- Stiggins, R. J., & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. *Educational and Psychological Measurement*, 2, 6–10, 26.
- Stroud, J. B. (1946). *Psychology in education*. Longmans, Green and Co.
- Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology*, 79(4), 535.
- Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of test development* (pp. 653–673). Lawrence Erlbaum.
- Trackman. (n.d.) *Play a better game*. <http://www.trackman.com>
- Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology*, 93(3), 711.
- U.S. Government Accountability Office. (2011). *Higher education and disability: Improved federal enforcement needed to better protect students' rights to testing accommodations*. (GAO 12-40). USGAO. <https://www.gao.gov/assets/gao-12-40.pdf>
- Valve anti-cheating software. (2020). In *Wikipedia*. https://en.wikipedia.org/wiki/Valve_Anti-Cheat
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- Van Moere, A. (2013). Raters and ratings. *The Companion to Language Assessment*, 3, 1358–1374.
- Vantage Learning. (2001). *A preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, December 4–9). Attention is all you need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)* (Vol 30, pp. 1–11) Curran Associates, Inc.

- Walker, A. A., & Engelhard, G. (2014). Game-based assessments: A promising way to create idiographic perspectives. *Measurement: Interdisciplinary Research and Perspectives*, 12(1–2), 57–61. <https://doi.org/10.1080/15366367.2014.921037>
- Wallach, H. M. (2006, Jun 25–29). Topic modeling: Beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 977–984), ACM.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GLUE: A multi-task benchmark and analysis platform for natural language understanding*. ArXiv <https://doi.org/10.18653/v1/W18-5446>.
- Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., & Huang, X. (2021, July 18–24). Unispeech: Unified speech representation learning with labeled and unlabeled data. *Proceedings of the 38th International Conference on Machine Learning* (pp. 10937–10947), Online Virtual.
- Wang, J., Engelhard, G., & Wolfe, E. W. (2016). Evaluating rater accuracy in rater-mediated assessments using an unfolding model. *Educational and Psychological Measurement*, 76(6), 1005–1025. <https://doi.org/10.1177/0013164415621606>
- Wang, S., Fellouris, G., & Chang, H.-H. (2019). Statistical foundations for computerized adaptive testing with response revision. *Psychometrika*, 84(2), 375–394. <https://doi.org/10.1007/s11336-019-09662-9>
- Wang, S., Zheng, Y., Zheng, C., Su, Y.-H., & Li, P. (2016). An automated test assembly design for a large-scale Chinese proficiency test. *Applied Psychological Measurement*, 40(3), 233–237. <https://doi.org/10.1177/0146621616628503>
- Wang, X., Evanini, K., Bruno, J., & Mulholland, M. (2016, December 13–16). Automatic plagiarism detection for spoken responses in an assessment of English language proficiency. *Proceedings of the 2016 IEEE Spoken Language Technology Workshop* (pp. 121–128), IEEE.
- Weiner, J. A., & Hartz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, 18(1), 13–20.
- Wheeler, J. M., Raczynski, K. R., Cohen, A. S., & Engelhard Jr, G. (2024). Using topic models to understand rater-mediated writing assessments. *The Journal of Experimental Education*, 92(1), 186–205.
- Wheeler, J. M., Wang, S., Tan, Y., & Cohen, A. S. (2022). *Textual data as process data: A new scoring procedure for mixed-format assessments*. National Council on Measurement in Education.
- Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1), 45–55. <https://doi.org/10.1177/016555159201800106>
- Wilcox, K. T., Jacobucci, R., & Zhang, Z. (2020). Bayesian supervised topic modeling with covariates. *Multivariate Behavioral Research*, 55(1), 141–141.
- Wild, F. (2016). Representing and analysing meaning with LSA. In F. Wild (Ed.), *Learning analytics in R with SNA, LSA, and MPIA* (pp. 71–106). Springer. https://doi.org/10.1007/978-3-319-28791-1_4

- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated essay scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13.
- Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, 43(2), 159–171.
- Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context*. Pearson. http://images.pearsonassessments.com/images/tmrs/tmrs_rg/RaterEffects_101510.pdf
- Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31–37. <https://doi.org/https://doi.org/10.1111/j.1745-3992.2012.00241.x>
- Wolfe, E. W., Moulder, B. C., Myford, C. M. (1999, April 19–23). *Detecting differential rater functioning over time (DRIFT) using a Rasch Model* [Paper presentation]. American Educational Research Association Annual Meeting, Quebec, Canada.
- Wollack, J. A., & Schoenig, R. W. (2018). Cheating. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 260–265). Sage.
- Wu, C., Wu, F., Wu, S., Yuan, Z., Liu, J., & Huang, Y. (2019). Semi-supervised dimensional sentiment analysis with variational autoencoder. *Knowledge-Based Systems*, 165, 30–39.
- Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech Using SpeechRater(SM) v1.0* (ETS Research Report No. RR-08-62). ETS.
- Xie, S., Evanini, K., & Zechner, K. (2012, June 3–8). Exploring content features for automated speech scoring. *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 103–111), Montréal, Canada. ACL.
- Xiong, H., Cheng, Y., Zhao, W., & Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. *Computers & Industrial Engineering*, 135, 333–347.
- Xiong, J., & Li, F. (2023). Bilevel topic model-based multitask learning for constructed-responses multidimensional automated scoring and interpretation. *Educational Measurement: Issues and Practice*, 42(2), 42–61. <https://doi.org/10.1111/emip.12550>
- Xiong, W., Wu, L., Allewa, F., Droppo, J., Huang, X., & Stolcke, A. (2018, April 15–20). *The Microsoft 2017 conversational speech recognition system*. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5934–5938), Calgary, AB, Canada. IEEE.
- Xu, W., & Liu, M. (2016). Using Coh-Metrix to analyze Chinese ESL learners' writing. *International Journal of Learning, Teaching, and Educational Research*, 15(5), 16–26.
- Yao, X. M. (2013). LSA-based automated essay scoring in Chinese context. *Applied Mechanics and Materials*, 274, 654–657. <https://doi.org/10.4028/www.scientific.net/amm.274.654>

- Yeung, S. F., & Lui, J. C. S. (2008). Dynamic Bayesian approach for detecting cheats in multi-player online games. *Multimedia Systems*, 14(4), 221–236. <https://doi.org/10.1007/s00530-008-0113-5>
- Yoon, S. Y., Bhat, S., & Zechner, K. (2012, June 3–8). Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 180–189), Montréal, Canada. ACL.
- Yoon, S. Y., Evanini, K., & Zechner, K. (2011, June 24). Non-scorable response detection for automated speaking proficiency assessment. *Proceedings of the sixth workshop on innovative use of NLP for Building Educational Applications* (pp. 152–160). ACL.
- Young, J. W., & King, T. C. (2008). *Testing accommodations for English language learners: A review of state and district policies* (Research Report 2008-6). College Board.
- Young, R. E. (1982). Concepts of art and the teaching of writing. In J. J. Murphy (Ed.), *The rhetorical tradition and modern writing* (pp. 130–141). Modern Language Association.
- Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10), 883–895. <http://www.sciencedirect.com/science/article/pii/S0167639309000703>
- Zechner, K., & Loukina, A. (2020). Automated scoring of extended spontaneous speech. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 365–382). Chapman & Hall/CRC. <https://doi.org/https://doi.org/10.1201/9781351264808>
- Zenisky, A. L., & Sireci, S. G. (2007). *A summary of the research on the effects of test accommodations: 2005–2006*. Report No. 47. University of Minnesota, National Center for Educational Outcomes.
- Zhai, S., & Zhang, Z. M. (2016, February 12–17). Semisupervised autoencoder for sentiment analysis. *Proceedings of the thirtieth AAAI Conference on Artificial Intelligence* (pp. 1394–1400), AAAI.
- Zu Eissen, S. M., & Stein, B. (2006, April 10–12). Intrinsic plagiarism detection. *Proceedings of the 28th European Conference on Advances in Information Retrieval* (pp. 565–569), London, United Kingdom. Springer-Verlag.

NOTES

1. The CEFR is used to evaluate other languages besides English.
2. Word error rate = (substitutions + insertions + deletions)/number of words spoken, where errors are:
 - substitution: when a word is replaced (for example, “shipping” is transcribed as “sipping”),
 - insertion: when a word is added that was not said (for example, “hostess” is transcribed as “host is”), and
 - deletion: when a word is omitted from the transcript (for example, “get it done” is transcribed as “get done”).

Technology- Based Assessment

Validity, Modeling, and Analysis Issues

Randy E. Bennett
ETS

Michelle LaMar
ETS (Retired)

John Mazzeo
ETS (Retired)

In this chapter, we focus on validity, modeling, and analysis issues in technology-based assessment. For present purposes, we define *technology-based assessment* (TBA) as a measurement used for decision-making primarily in education, but also in the workplace, that employs digital computing in most, if not all, aspects of its creation, delivery, presentation, scoring, or reporting.

As of this writing, TBAs were being used for consequential decision-making purposes widely in the United States, as well as in some other countries and in some international assessments. Among the more prominent of such programs are the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), the Australian National Assessment Program, the GRE, the TOEFL, the Graduate Management Admission Test, the United States Medical Licensing Examination (USMLE), the Law School Admission Test, and the California Assessment of Student Performance and Progress (CAASPP). CAASPP is particularly notable because of its test-taker volume, which may be the largest of any such measure. On a single day, June 7, 2019, that volume exceeded 670,000 students, with well over 3 million individuals taking the examinations in the spring 2019 cycle (Johnson, 2019). Although many testing programs, especially those outside the United States, continue to test on paper, the number of major testing programs that have become digital, coupled with the large size of CAASPP, demonstrates that technology delivery is feasible at scale. The transition has proven to be, as one of the authors of this chapter had earlier suggested, “inevitable and inexorable” (Bennett, 2002).

What is motivating this transition? Briefly stated, the most salient reasons revolve around three areas. First is the need to align the medium of testing with that of learning and of the information economy’s workplace. Absent that alignment, assessment runs the risk of appearing, and becoming, irrelevant to its constituents. The second reason is the belief that assessment processes can be conducted more efficiently, with savings in time and cost. Scores, for example, can in some instances be generated and reported immediately. Third is the measurement of constructs that are impossible to evaluate in traditional testing modes. Examples include using technology tools for such activities as reading in hyperlinked environments, information search and synthesis, writing, modeling, and collaborative problem-solving (Institute for Education Sciences, n.d., 2018; Mullis & Prendergast, 2017).

The transition that has occurred since the 1990s in the United States and elsewhere can be described at a high level in terms of three stages or generations (Bennett, 1998, 2010a). The first generation is essentially an infrastructure-building effort. The cost and time required to put that infrastructure into place are typically substantial, involving obtaining hardware and software, hiring personnel, providing training, and creating myriad new processes and procedures. To control cost and complication, tests in this generation often look little different from paper assessments in their realized design and question format. This generation’s tests primarily serve institutional purposes like school accountability, leverage technology minimally (e.g., through incremental advances like adaptive testing), and are generally organized as singular events (e.g., for state accountability purposes, as annual, end-of-year administrations).

In the second generation, goals begin to shift toward achieving some degree of qualitative change and to leveraging the technology more effectively for efficiency improvement (Bennett, 1998, 2010a). With respect to the former goal, a diversity of less traditional item formats may be employed, including those sometimes called “technology-enhanced” (e.g., involving multimedia stimuli, dragging and dropping onscreen text or objects, highlighting segments of text). Additionally, new constructs may be introduced (e.g., writing on computer), so that what is assessed begins to change fundamentally. Efficiency improvements extend beyond delivery via automated approaches to item generation (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2010), test assembly (Veldkamp & Paap, 2017), and scoring (Shermis & Burstein, 2013; see also Shermis et al., this volume), as well as to using the Internet for such processes as item review, standard setting, human online scoring, reporting, and other communications with test users.

Reinvention characterizes the third generation of assessments (Bennett, 1998, 2010a). What in the two earlier generations was an evolution dictated primarily by technology now shifts to one driven by substance. In this generation, theory-based models and cognitive principles combine with more traditional content considerations to provide the substantive basis for assessment design (see Huff et al., this volume). Second, these assessments integrate the needs of individuals more fully with those of institutions. A special case of this development is greater integration with instruction, including the repeated sampling of performance over time. Third and finally, the use of complex simulations and other interactive performance tasks allows new skills to be measured and traditional ones to be evaluated in more meaningful ways.

Of note with respect to the focus of this chapter is that the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association [AERA] et al., 2014) contains no section or set of standards specifically devoted to technology-based assessment. The document does, however, make clear that “computer . . . tests need to be held to the same requirements of technical quality as other tests” (p. 197) and that the “interpretation of scores on technology-based tests are evaluated by the same standards for validity, reliability/precision, and fairness as tests administered through more traditional means” (p. 188). In this sense, an assessment’s purpose, with its attendant claims for score interpretation and use, remains the central consideration in determining the appropriate level of technical quality.

With the above as an introduction, the goals of this chapter are to present (a) major issues in validity, modeling, and analysis for TBAs and (b) potential approaches to addressing those issues. The chapter is organized as follows. The first section centers on assessments used to support consequential purposes. This class encompasses decisions that, as singular instances, may have highly significant impact on individuals, groups, or institutions and that are often difficult to reverse. Examples include school admissions, promotion and graduation, educator evaluation, school accountability, intranational and international comparisons, and job licensure and certification. In this section, the history and current landscape are described for technology-based tests of this type.

Discussed are innovative tasks and item types, comparability, the analysis of response processes, and the automated scoring of complex constructed-response tasks.

The second major section covers assessments employed for in-the-moment instructional decisions or for describing what a student knows and can do so that near-term (but not necessarily real-time) instructional next steps can be taken. As singular instances, these decisions usually have less dramatic impact on individuals and are more easily reversed than decisions we have termed *consequential*. This section also begins with a review of the history and current landscape that includes brief consideration of common technology-based assessment designs and task types, response processes, and scoring methods. This section closes with a discussion of psychometric models for assessment embedded in instructional and learning systems.

The chapter's last major section explores the third-generation idea of combining both assessment purposes—that is, consequential decision-making and instructional support—in the same assessment. We conclude the chapter by summarizing key points, giving recommendations for research, and speculating on future directions.

ASSESSMENTS USED TO SUPPORT CONSEQUENTIAL PURPOSES

In this section, we outline the landscape regarding TBAs used for consequential purposes in education, as well as in the professions and occupations. We include tests used to make consequential decisions about individual test takers, as well as state, federal, and international measures intended to provide results for various monitoring and accountability purposes. In doing so, we trace major milestones in the development and operational deployment of TBAs.

Even prior to the widespread availability of desktop computers, the potential benefits of testing by computer were relatively obvious. Such delivery offered the possibility of immediate scoring and reductions in cost associated with the elimination of printing and shipping of test materials. Throughout the 1970s and into the 1980s, advances in psychometric research and theory by Lord (1970) and Weiss (1976) revealed how more powerful psychometric models, coupled with alternative test delivery schemes enabled by computer, could also improve quality. Computer delivery that moved beyond simple, linear, preassembled test forms to the use of sequential and computer-adaptive tests (CAT) offered the potential of increased measurement precision throughout the proficiency range, as well as increased efficiency (i.e., better precision for a fixed amount of testing time; Lord, 1980; Weiss, 1982).

The operationalization of TBA required an infrastructure to deliver tests securely on a large scale. As described by W. D. Way and Robin (2016), this infrastructure first emerged, surprisingly, from the University of Illinois's efforts to expand computer-assisted instruction (CAI) as implemented in its Programmed Logic for Automatic Teacher Operations system. That system was further developed and commercialized by Control Data Corporation. In partnership with the National

Association of Securities Dealers, the Control Data Corporation used this system, circa 1978, to offer what was most likely the first proctored TBA for consequential decisions, in this instance for licensure and certification.

One of the earliest entities interested in the potential convenience and efficiency of TBA—in particular, through adaptive testing—was the U.S. military (Sands et al., 1997). Military institutions sponsored considerable research and influential conferences on this topic that advanced the measurement field substantially (Weiss, 1978, 1980, 1985). In 1979, a feasibility study was launched to develop an adaptive version of the military's selection and placement test battery, the Armed Services Vocational Aptitude Battery. This test was developed, evaluated, launched for operational testing in 1992, and adopted fully in 1996.

One of the earliest efforts to deploy TBAs for consequential decisions in education—in this case, college placement decisions—began in the mid-1980s, when the College Board introduced its computerized-adaptive College Placement Tests (CPTs). The CPTs, delivered on microcomputers, were intended for use at 2- and 4-year institutions to help identify incoming students requiring remedial education in English, reading, and mathematics. The development and validation of these tests is described by Ward (1988) and Ward et al. (1986). The College Board's current college placement test—Accuplacer—is a direct descendant of the CPTs.

A subsequent major milestone was the development and introduction by ETS (Educational Testing Service) in 1993 of a CAT version of the GRE General Test, for use in graduate admissions decisions (Mills, 1999). Designed to award scores that were comparable to its paper-and-pencil predecessor, this work produced a key innovation: an item-selection algorithm capable of balancing psychometric considerations with constraints related to item content, format, and exposure, including dealing with passage-based items (Stocking & Swanson, 1993). This innovation helped to ensure that GRE CAT scores not only were psychometrically efficient but also met the content validity and fairness criteria important to consequential use. In this same decade, ETS followed with computer versions for other postsecondary admissions programs, including the adaptive Graduate Management Admission Test and the TOEFL CBT, with both adaptive and linear sections.

Whereas the introduction of these TBAs was a notable advance, it also surfaced some unanticipated challenges (Wainer & Eignor, 2000). Paper-and-pencil tests could be administered to large numbers of students on a few days per year, requiring only a small number of different forms. In contrast, because of infrastructure limits, TBAs had to be delivered in small centers to fewer students on a more continuous basis. Since demand tended to peak at particular times, problems quickly arose with test-taker access to centers. As important, it soon became apparent that initial plans for item pool sizes and item exposure controls were inadequate to maintain security. Dealing effectively with these problems substantially increased the costs of CAT TBA relative to paper-and-pencil administration. Additional problems emerged with respect to preventing students from learning how to “game” the CAT (W. D. Way & Robin, 2016). The vast

majority of these problems have been ameliorated over time, with increased access, use of alternative models such as multistage adaptive testing (MST), and enhanced schemes for exposure control. As of this writing, the two major undergraduate college admissions programs (ACT, formerly American College Testing; and SAT, formerly Scholastic Aptitude Test) have begun to offer TBA versions, with the SAT having become all digital as of 2024 (College Board, 2024).

The first forays into TBA consisted largely of adaptively delivering multiple-choice items, the staple of paper-and-pencil testing programs. Relatively little was done to leverage the affordances of TBAs for presenting more complex tasks focused on measuring a broader set of competencies. While this eventuality may have been partly due to the limitations of the available computer technology, other contributing factors included cost, limits on available testing time, score reliability, comparability with paper versions, and fairness concerning the use of performance testing generally.

It is perhaps not surprising, then, that early attempts to introduce performance tasks occurred with licensure and certification testing, where longer testing times, costs that could be passed on to the test taker, and reduced need to maintain comparability with prior versions made such innovations more tractable. For example, in the 1990s the National Council of Architectural Registration Boards introduced an automatically scored, performance-based design section into its licensure examination (Bejar & Braun, 1999). Similarly, the National Board of Medical Examiners incorporated computer-based patient management cases into the USMLE (Margolis & Clauser, 2006). Finally, the American Institutes of Certified Public Accountants began to administer TBAs containing performance tasks (Breithaupt et al., 2006). It is noteworthy that all these examinations retained very substantial, complementary multiple-choice sections to achieve the reliability and generalizability levels needed for licensure decisions. (See Margolis et al., this volume, for a comprehensive review of testing in licensure and certification.)

The first waves of consequential TBA testing—in the military, in university placement, in graduate admissions, and in licensure and certification—occurred with TBAs delivered on mainframes, then on stand-alone desktop computers in college placement offices, and next via networked machines at vendor testing centers (e.g., Prometric and Pearson VUE) and in university labs. Hardware and software could be reasonably standardized, and the use of data networks to transmit test content and response data allowed security to be maintained. For K–12 education, in contrast, the use of TBA for state-mandated accountability tests, high school end-of-course measures, and graduation examinations awaited development of a more pervasive infrastructure that could accommodate machines resident in the schools. (See Ho & Polikoff, this volume, for a comprehensive discussion of assessment for accountability in K–12.)

With the evolution of the Internet and the proliferation of laptops and tablets, TBA became increasingly practical for state-mandated accountability tests. This history dates to about 2000, when Oregon, Virginia, and a few other states each began pilot TBA programs (Bennett, 2002). Those pilot efforts gradually expanded so that by

2007, Oregon, for example, had received approval for a computer-adaptive version of its federally mandated K–8 summative assessment (W. D. Way & Robin, 2016). The passage of the American Recovery and Reinvestment Act of 2009, and in particular the Race to the Top Assessment Program (U.S. Department of Education, n.d.), gave a dramatic boost to development by funding several testing consortia, including the Smarter Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Careers (PARCC). Smarter Balanced and PARCC developed TBAs aligned to the Common Core State Standards in English language arts and mathematics. The tests made use of adaptivity (Smarter Balanced), items that employed technology affordances (technology-enhanced items [TEIs]), and more innovative performance tasks to measure writing, research, and problem-solving skills. The tests were first administered operationally in 2015 in a substantial number of states. By the 2015–2016 school year, between the adoption of the consortia tests and states that had implemented their own TBAs, EdTech Strategies (2015) estimated that 85% of the accountability tests in Grades 3 to 8 would be delivered online. Moreover, although many states were using both online and paper-based tests, only three states were not using some form of TBA. Primarily a result of political pressures, state membership in the consortia has waned substantially, with many states since choosing to implement their own TBA programs. Despite the reduction in consortia membership, most states continue to use TBA in whole or in part for their state-mandated accountability assessments (Olson, 2019).

Progressing alongside the development of TBA for state accountability was incorporation into the U.S. national assessment. In 1999, NAEP began a series of studies to facilitate its transition to TBA. NAEP is a congressionally mandated assessment and differs from typical testing programs in that group-level results, rather than scores for individual students, are reported for the nation, states, selected demographic groups, and some large-city school districts. Results are based on samples of schools and students, with a highly efficient matrix-sampling design employed to cover a broad content domain while minimizing testing time for any individual. Although the results do not carry direct consequences for schools or students, NAEP results receive much press coverage and are highly influential in shaping national, state, and big-city education policy.

The NAEP TBA studies investigated the delivery of traditional NAEP paper-based assessments in mathematics and in writing via computer (Bennett et al., 2008; Horkay et al., 2006; Sandene et al., 2005). Computer delivery occurred on desktop computers via the Internet or on disconnected laptops. These studies, which involved administering the assessments in both modes to randomly equivalent samples, examined issues related to comparability of results between computer and paper administration and between laptop and desktop administration, as well as issues related to differential subgroup performance and the role of computer familiarity. Results were largely favorable, suggesting the prospect of transitioning NAEP to TBA while maintaining trend comparisons to previous paper results. A third study explored the use of innovative interactive tasks

to examine problem-solving in technology-rich environments (Bennett et al., 2007, 2010). This study also yielded promising results.

Based on this early work, subsequent research, and developments in technology and testing, as of this writing, the NAEP program has almost completely replaced its paper-based assessments with TBAs. In 2011, NAEP conducted its first operational writing TBA in Grades 8 and 12 (National Center for Education Statistics, 2012). The assessment, which was delivered on NAEP laptops, required students to compose essays in response to prompts, just as in earlier paper administrations. However, because the assessment was based on a new framework, no attempt was made to maintain trends to prior years. In 2014, NAEP conducted the Technology and Engineering Literacy assessment, designed from inception as a TBA (National Center for Educational Statistics, n.d.). This assessment included TEIs as well as scenario-based interactive tasks (see the section on “Innovative Item, Task, and Assessment Types”).

NAEP’s mathematics and reading assessments at Grades 4 and 8 were transitioned in 2017 and consisted largely of traditional multiple-choice and constructed-response items, along with a limited number of TEIs (Jewsbury et al., 2020). Plans for future assessments in all NAEP subjects presume digital delivery and the increased use of innovative item types. While the current TBAs still employ the standard NAEP matrix-sample designs, adaptive administration has also been explored (Oranje et al., 2014).

In similar fashion to NAEP, such international group-score assessments as PISA and the Trends in International Mathematics and Science Study (TIMSS), are transitioning or have transitioned to TBA. (See Braun & Kirsch, this volume, for a comprehensive discussion of assessments in the international context.) Unlike NAEP, these assessments continue to support both TBA and paper administration because not all participating education systems have the required infrastructure. Consequently, these programs work to maintain comparability of results, both to prior assessment cycles and within a cycle.

PISA, which measures the reading, mathematics, and science skills of 15-year-olds across many nations and jurisdictions, transitioned to TBA for all three content areas (plus financial literacy) in 2015 (Organisation for Economic Co-operation and Development [OECD], 2017). The assessments, which were delivered on school desktops and laptops, consisted largely of TBA analogues to existing paper item types. In 2018, the reading TBA was converted to an MST, whereas the mathematics and science assessments remained linear tests.

TIMSS, which assesses the mathematics and science skills of eighth graders, was also in the process of transitioning to a digitally based version (Mullis, 2019). As of 2019, TIMSS was administered via computers or tablets, as well as in its paper format, with about half of participating education systems utilizing each mode. In addition to traditional items, the eTIMSS digital version included innovative tasks designed to simulate real-world and laboratory situations where students can integrate and apply processes and content knowledge to solve mathematics problems and conduct scientific experiments.

A third international group-score assessment, the Progress in International Reading Literacy Study (PIRLS), has also incorporated a TBA, this one concentrating on the reading skills of fourth graders. In 2016, 16 of the 58 education systems that participated in the paper-based assessment (including the United States), also took ePIRLS, a supplementary measure of online reading skills (Institute for Education Sciences, n.d.; Mullis & Prendergast, 2017). The assessment, which was delivered on either school desktop or laptop computers, consisted of two tasks that required test takers to answer multiple-choice and constructed-response items pertaining to information presented on simulated web pages.

As we look to the future, it seems reasonable to assume that TBAs will move toward a more decentralized delivery model admitting a wider range of assessment devices. As seen with the consortia-developed K–12 tests, beyond requiring a minimum configuration, it is impractical to impose strict control over the delivery device. Within and across U.S. states, schools own a wide variety of equipment. Furthermore, it is not feasible for test sponsors to provide a common device at that scale. Similar considerations exist for international programs like PISA and TIMSS.

Moreover, disruptive events like the COVID-19 pandemic make it evident that consequential testing may need to be carried out remotely at times. In pandemic conditions, gathering in groups is unsafe. Thus, consequential testing may increasingly need to be administered in homes on test taker–owned equipment, where feasible and where the security conditions for such testing are acceptable to both sponsors and users. For the purposes of K–12 assessment, approaches such as remote proctoring involving live monitoring or video capture, as well as recording and analyzing process data (e.g., timing and keystroke sequences), remain controversial as of this writing.

In addition to various exogenous factors, validity and fairness concerns may be moving consequential testing toward a more decentralized delivery model. As more complex constructs are included and tasks are incorporated that require extensive use of and familiarity with the device and interface, issues of comparability and fairness might be better served by allowing test takers to work on their own equipment. In that way, score differences might be more likely to reflect differences in the target constructs rather than also reflecting familiarity with an arbitrarily imposed, standardized device configuration.

To make consequential testing possible in decentralized testing environments, current strategies for ensuring standardization, security, and fairness will need to be adapted. As for K–12 accountability testing, some control over testing conditions may still be imposed through the specification of minimum device characteristics like screen size and resolution, types of keyboards, and transmission bandwidths. Providing free online practice materials and tutorials to familiarize test takers with task types, tools, and interfaces will also continue to be important. But most likely, these strategies will need to be coupled with a careful approach to designing delivery systems, interfaces, tools, and tasks to ensure performance is device agnostic.

Innovative Item, Task, and Assessment Types

In the preceding section, we outlined the development and landscape for consequential TBA. Next, we consider the building blocks of innovation in such assessments—items, tasks, and assessment types.

Among the potential benefits technology brings to assessment design are increased construct fidelity (Russell, 2016), evoking and supporting more complex cognitive processes, and allowing greater observability of the response process. Construct fidelity represents the extent to which the response process evoked by the item accurately reflects the targeted measurement construct. Russell and Moncaleano (2019) identified two components of construct fidelity: the extent to which the problem context is consonant with domain practice and the degree to which the response interaction aligns with the interactions that would occur in a real-world situation. For most constructs, paper-and-pencil items are far from real-world applications in both respects, implying that technology enhancement might allow for increases in construct fidelity.

Complexity of cognitive process relates to the length and depth of reasoning that is required to correctly respond to the task. Criticism has long been made about traditional multiple-choice item formats because they can measure only certain types of cognition and because of the impact that focus may have on teaching and learning (Frederiksen, 1984). Technology-enabled assessments have the potential to measure constructs that involve more complex cognitive processes, such as science practices, collaboration skills, and investigative research (Bennett et al., 2010; Csapo et al., 2012). Through use of simulated problem contexts, computer-based tasks can support an extended problem-solving process more appropriate for these domains from which student cognition might be modeled at a finer grain size (Mislevy & DiCerbo, 2012).

Whereas TBA may be useful in evoking complex cognition, traditional paper-based assessments can also call on higher order response processes through the use of text and diagrams (Bennett, 2012). An advantage of computer-based assessments, however, is the ease with which evidence about response process can be captured. The more interactive an item response, the more process data can be recorded. Thus, technology makes complex response processes more observable. “Observable” does not necessarily mean scorable; as discussed in the section “Response Processes,” challenges in interpreting and scoring process data must also be overcome before this evidence can be used meaningfully.

In the following paragraphs, we examine technological innovations in four broad categories: TEIs, extended-interaction items, scenario-based tasks, and simulation-based performance tasks. The first two categories involve innovations at the item level, while the second two encompass innovations for lengthier tasks that may include interdependent components.

TEIs

The most common items used in computer-based assessments are traditional item types with constrained outcome spaces, some of which could have an exact paper-and-pencil replica. These items might take multiple-choice, matching, sorting, or short-answer

formats. They may appear as discrete samplings, independent of one another, or as part of an item set. For assessments offered on both paper and computer, items with exact or highly similar counterparts are desirable to increase the comparability of score meaning between delivery modes. Additionally, these items produce data that fit commonly used scoring and psychometric models, increasing the likelihood that results can be put on the same scale as previous paper administrations.

While the format of such items is basically traditional, technical innovations can be introduced in stimulus presentation or response process, thereby making the item a “technology-enhanced” one. For example, the use of images, audio clips, animation, and video enables a richer presentation of information, which may increase construct fidelity or evoke more complex cognitive processes, even when the response type is highly constrained (Bennett et al., 1999). Figure 9.1 gives an early example that shows how multimedia might be employed in an item stimulus as part of a test of U.S. history knowledge. The item allows the student to view and analyze a primary source in a way closer to that of a historian.

For the response process, the affordances of mouse and touchscreen technology are frequently used to enable drag-and-drop or on-screen drawing as response mechanisms (Zenisky & Sireci, 2002). Drag and drop can be used for questions that ask test takers to move objects into the correct order or for items in which test takers reposition

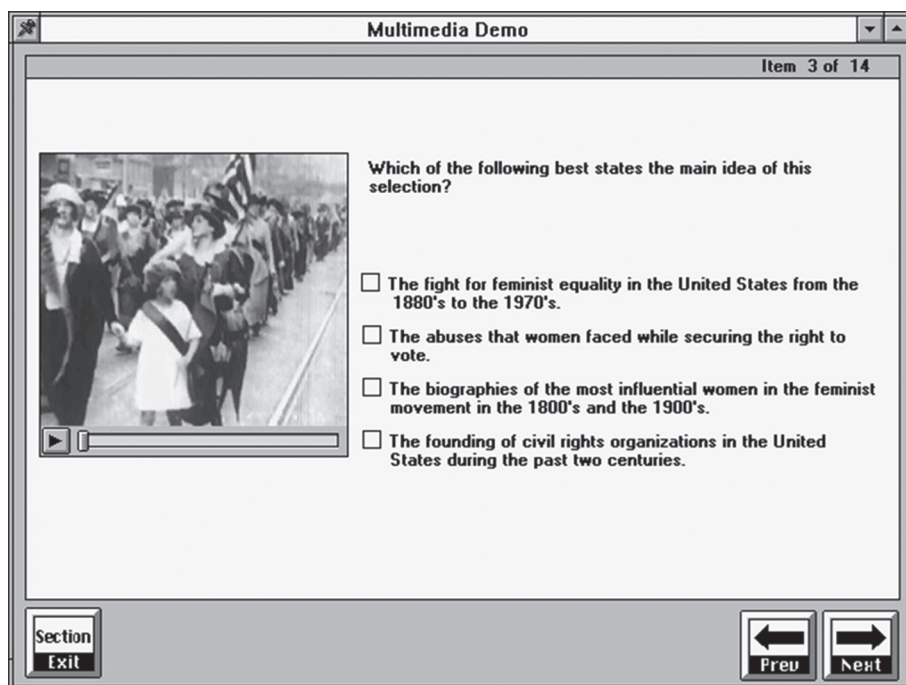


FIGURE 9.1

A Technology Enhanced U.S. History Item Using Multimedia

Note. Copyright ETS® 1999. Reprinted with permission.

objects from one set to match objects in another set. Additionally, drag and drop can access a more continuous response space, such as putting objects onto a map or placing a marker on a spectrum (ATP, 2017; Wan & Henley, 2012).

An example is the PISA 2015 interactive science task, *Fish Farm*. In this item set, students are given a brief context statement about the need to develop a particular type of alternative seafood source and the challenges encountered in operating such a source. Following this context are three independent items (i.e., subsequent items do not build directly on preceding ones or lead to a culminating task satisfying a goal given in the context statement). One of the set's items asks students to help design a sustainable farm ecosystem by dragging organisms into the appropriate tanks (Figure 9.2). The cognitive complexity of this item is fairly high because it “requires students to understand a system and the role of several organisms within that system” (OECD, 2018). Whereas this item could be rendered on paper using a diagram and labels for the organisms, the ability to move the organisms into the fish farm allows test takers to more naturally construct, evaluate, and revise their model.

Russell and Moncaleano (2019) judged that most usage of drag and drop in TBA failed to increase construct fidelity. Interactivity, per se, does not necessarily improve the item. In the fish tank task, the graphical representation allows students to reason

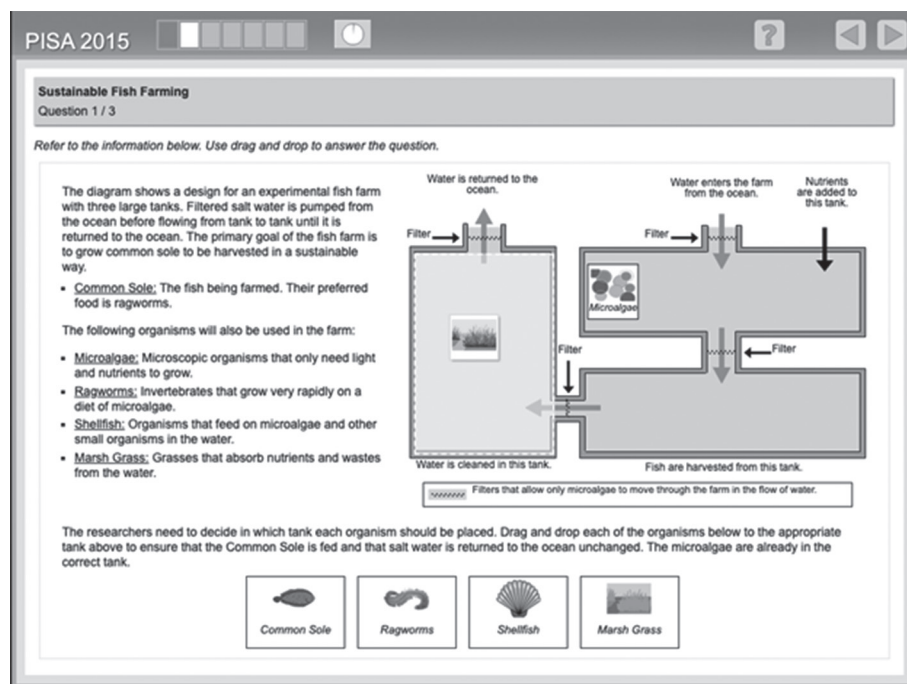


FIGURE 9.2

PISA 2015 Interactive Science Item in Which Students Use Drag-and-Drop Interactions to Complete a Model of a Fish Farm

Note. PISA = Programme for International Student Assessment. From *Try PISA 2015 Test Questions* by OECD, 2018. (<https://www.oecd.org/pisa/test/pisa2015/#d.en.537240>). CC by 3.0

more directly about the placement and interactions among the organisms. Thus, the drag-and-drop functionality is used effectively to increase the similarity of the evoked cognitive process to ones that a scientist might employ, arguably enhancing construct fidelity.

Extended-Interaction Items

More innovative uses of technology have generally resulted in greater interactivity, providing support for and observation of a more extended response process but still within the context of a discrete, stand-alone item or item set. These items frequently involve constructing a complex response to solve a problem. Test takers might be asked to graphically represent a process, analyze data, or test a hypothesis using a runnable model (a function that accepts input parameters and outputs a result intended to mimic reality, as in the “running in hot weather” example that follows). This type of item is distinguished both by the multiple actions a test taker is expected to take and the depth of cognitive processing associated with those actions.

An example of this class is shown in Figure 9.3. In this PISA 2015 item set, test takers are asked to use a runnable model to answer questions about factors that increase the risk of dehydration or heat stroke for a jogger on a hot day. Test takers are expected to interact with the model, running several trials with different settings before responding

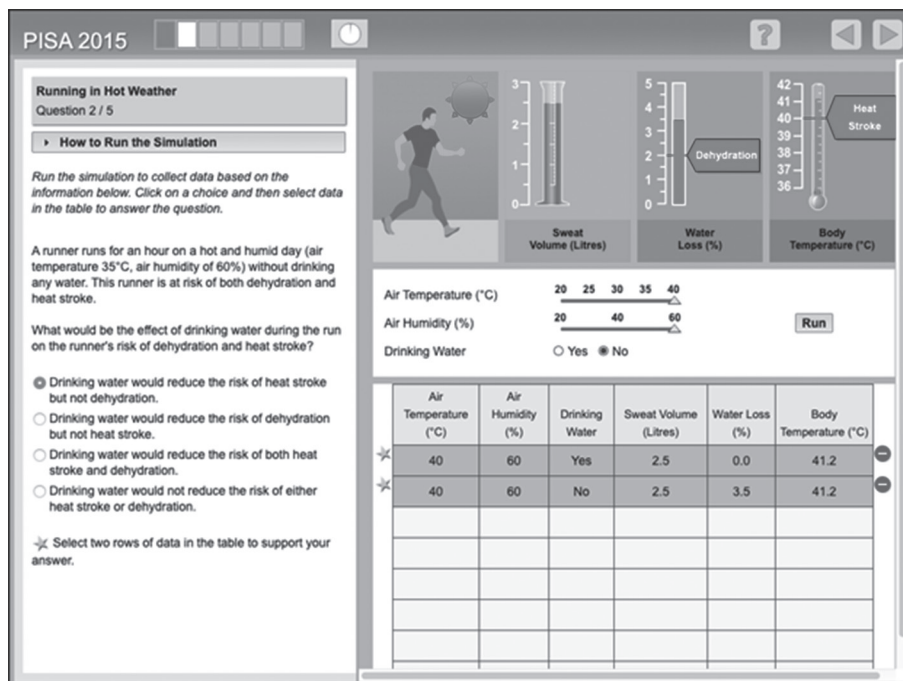


FIGURE 9.3

PISA 2015 Extended-Interaction Item From the *Running in Hot Weather* Set

Note. PISA = Programme for International Student Assessment. From *Try PISA 2015 Test Questions* by OECD, 2018. (<https://www.oecd.org/pisa/test/pisa2015/#d.en.537240>). CC by 3.0

to the multiple-choice item prompt. The interactions allow students to engage with a real-world problem, while providing evidence of solution process through event logs (discussed in the section “Types of Response Process Data”). Thus, this format permits measurement of a construct that is not possible to assess with a paper test.

Another example of extended interactivity can be found in the PISA 2018 literacy items, which provide a simulated web browser for test takers to employ in researching a topic. While the primary construct is reading comprehension, the item design requires students to follow hyperlinks and use tabs for navigation to relevant content. In this case, the interactivity is intended to increase construct fidelity by replicating the context in which one might undertake information search.

Scenario-Based Tasks

Whereas TEIs and extended-interaction items can provide richer stimuli and the opportunity to observe response processes, their design as independent items (including in sets) limits the depth and complexity of problem-solving. To tap deeper problem-solving processes, assessment designers may turn to scenario-based tasks (SBTs), as defined by Deane et al. (2018), O'Reilly et al. (2019), and others. These tasks are characterized by an overarching narrative, or scenario, which poses a goal for the test taker to achieve. The scenario presentation is followed by a sequence of related technology-enhanced and more traditional items that lead to a culminating performance in which the test taker attempts to satisfy the goal (e.g., a proposal for how a school might use a generous gift, backed by reasons and evidence from given sources). SBTs break the process leading to this goal into steps, each of which is at the same time part of the larger whole but implemented as one or more distinct items.

Some of the tasks included in the NAEP 2014 Technology and Engineering Literacy assessment offer examples. As an instance, the bike lane task consists of five items (see Figure 9.4). Students are first introduced to the motivating problem of safety when riding in a lane adjacent to automobile traffic. Initial items ask students to interact with a runnable road-sharing model that rates bike lane safety based on the manipulable parameters of automobile speed limit and lane width. In later items, students use the information gathered to create safe road designs for cyclists and ultimately to reach the goal of a bike route that optimizes safety, cost, and route length.

The SBT format, when used to measure problem-solving processes, is a compromise between the discrete items of traditional assessment and the open-ended problem-solving characteristic of an extended project. When assessing complex problem-solving, the ability to gather evidence on the process competes with the level of constraint imposed on the problem space. If we allow the student full freedom within the problem space, it is more difficult to detect, understand, and score the problem-solving process used. In unconstrained tasks, the most reliable evidence of competency is typically the outcome. Unfortunately, the outcome may offer little information with which to distinguish and subsequently guide learners within the middle and lower regions of the proficiency distribution. When the task outcome is complex enough to show

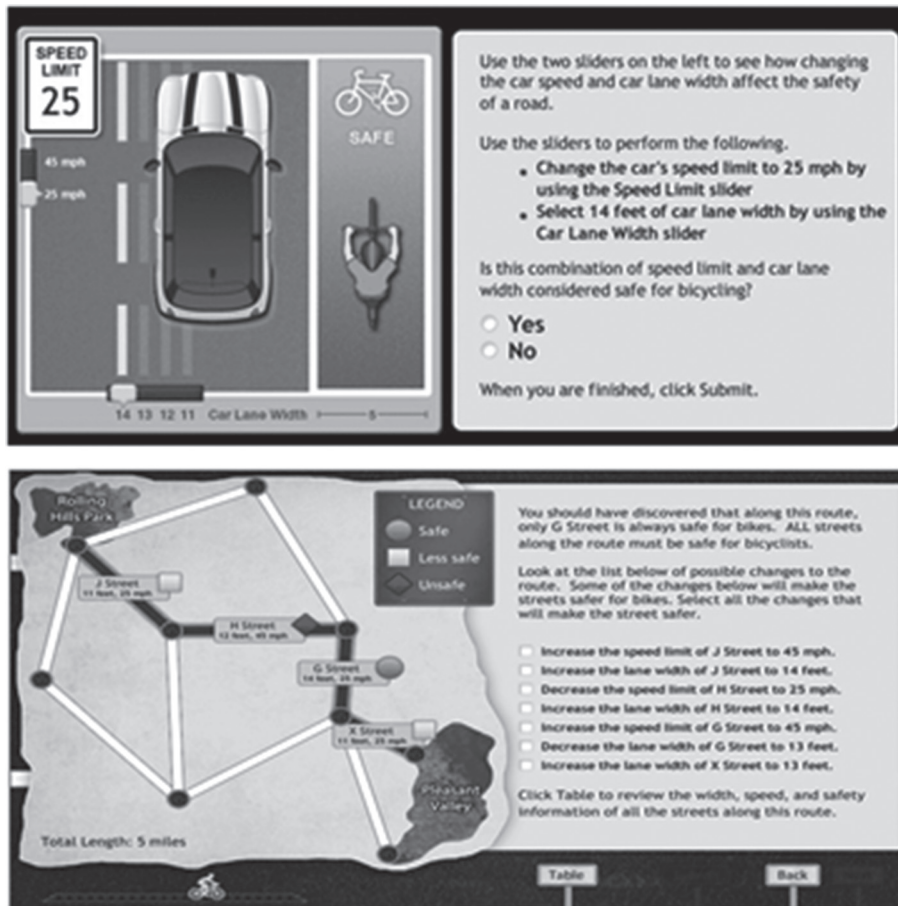


FIGURE 9.4
Selected Screens From the NAEP TEL 2014 Scenario-Based Task *Bike Lane*

Note. NAEP = National Assessment of Educational Progress; TEL = Technology and Engineering Literacy Assessment. From *2014 Technology and Engineering Literacy (TEL): Sample Scenario-Based Tasks*, by National Center for Education Statistics, n.d. https://www.nationsreportcard.gov/tel_2014/#tasks/bikelanes

finer gradations of competence, for example, as in an essay or lab report, making such distinctions comes at the cost of greater scoring time and effort.

If we instead discretize the problem-solving steps to allow observation of specific portions of the process, we are inevitably scaffolding problem-solving along a more constrained, and frequently more linear, path. That constraint could improve measurement or, conceivably, undermine it by giving the test taker aids not typically available in the criterion situation. Research suggests that the former situation may be the case. That is, SBTs may in some respects produce better measurement than less-structured performance tasks, particularly in more effectively aligning the processes measured with the intended construct (Guo et al. 2019, 2020; Zhang, Deane, et al., 2019). This result could be due to the initial questions in an SBT helping to activate relevant knowledge and better orient the test taker to the task at hand. In the criterion situation,

similar orientation can come, for example, from consulting with others and viewing examples of quality solutions to comparable problems.

Simulation-Enabled Performance Tasks

While SBTs scaffold the problem-solving process to enable observable evidence about the complex cognition underlying a performance, these tasks are somewhat unnatural. According to Russell's definition of construct fidelity (Russell & Moncaleano, 2019), SBTs may provide a realistic context, but because of their structure, the actions students take might diverge from those that a domain practitioner would employ in their problem-solving.

Performance assessments come closer to actual domain practice, allowing individuals to demonstrate a skill by carrying out a less structured task. A driving test offers a good example. However, although it has high construct fidelity, it requires specialized equipment (i.e., a car), a human examiner, and an individualized administration. A compromise position is offered by TBAs that employ simulation-enabled performance tasks. Such tasks attempt to represent key stimulus features of a given performance situation, calling on competencies that would be employed in that situation, but reduce the resource requirements to more manageable levels.

As an example, the USMLE seeks to assess a potential physician's ability to independently diagnose and treat a variety of patient conditions. As noted, in 1999, the National Board of Medical Examiners incorporated patient case management simulations into the computer-based USMLE (Dillon & Clauser, 2009). In each case management task, the patient's condition changes over simulated time. The test taker can engage in various actions that a doctor might take, including requesting patient history, getting the results of a physical exam, ordering laboratory tests, making a diagnosis, and prescribing treatment. While the "physical exam" is implemented using a set of checkboxes, the lab tests and treatment plan are specified using text input. This format prevents any form of prompting because the test taker will not see a list of available tests or treatment options (unless there are multiple options that match the test taker's input). While these simulations are not similar in fidelity to interacting with a real patient (they do not include any visual representation of the patient or of test results like X-rays), they replicate the main aspects of the problem-solving process used in practice. As mentioned, simulation-enabled performance tasks have also been employed in the architect and accountant examinations, among others (Clauser et al., 2016). (See Margolis et al., this volume, for more on assessment in licensure and certification.)

Less comprehensive simulation-enabled tasks can be found in school testing programs. For example, PISA 2015 assessed collaboration skills using a simulated group assignment in which the test taker negotiates with two automated agents to achieve a specified goal (OECD, 2018). In this instance, all choices made by the test taker were selected responses; however, the choices made changed the task situation.

Of special note in simulation-based performance tasks is that technology plays a key role in observing and making inferences about complex cognitive processes. That is,

the technology environment provides a reasonably realistic problem setting that allows the test taker to work through the problem in a high-fidelity manner. The test taker's step-by-step decisions can be recorded in computer log files, making inferences about their problem-solving process feasible and, in principle, scorable.

Appropriate Use of Innovative Items and Tasks

While innovative item types have important advantages, they are generally expensive to build, more complicated to administer, and less well understood in terms of measurement properties. For consequential testing, such item types should be used only when substantively appropriate, suited to the target population, and logistically feasible.

With respect to substantive appropriateness, as noted, such innovation can improve construct fidelity by enhancing problem presentation or response interaction, thereby evoking more relevant problem-solving processes. Instances include constructs that involve the use of technology itself, such as technical literacy or data analysis, as well as constructs for which the technology can provide a more faithful context for the application of skills, as in scientific inquiry, medical licensure, or remote collaboration (Bennett et al., 2010; Clauser et al., 2016; Csapo et al., 2012).

In addition to fit with the intended construct, appropriate use implies a good fit with the target population. Fit in this context means that all test takers can interact with items in ways that provide evidence of competency. Essential to this premise is an infrastructure capable of validly assessing *all* members of the test-taker population, including those with disabilities or who are English learners (see also Rodriguez & Thurlow, this volume).

Feasibility encompasses assessment design, development, administration, scoring, and reporting. Innovative items can introduce complications in any or all these phases. Design and development time and costs are frequently underestimated by staff members more familiar with traditional assessments. However, the costs tend to be front-loaded because the deployment of innovative item types requires developing or adopting new tools and processes, training assessment developers, designing for accessibility, conducting cognitive labs to evaluate whether the intended processes are evoked, etc. Once this foundation is established, the operational development of innovative items can be streamlined by using appropriate authoring tools and improved through the telemetry (i.e., process data) produced by the items. Finally, scoring and reporting requires appropriate psychometric models to facilitate defensible inferences.

Validity, Modeling, and Analysis Issues

Innovative items and tasks present new challenges for validity, modeling, and analysis, especially for tests used to make consequential decisions. Such innovations demand evidence that new presentation formats and response interactions produce meaningful inferences about the intended constructs. Studies may cover the range of evidence types recommended for validation generally (AERA et al., 2014), but with added attention to whether the response processes evoked are consistent with the target construct

and to what degree construct-irrelevant variance might have been introduced (e.g., Gallagher et al., 2002; Mislevy et al., 1999). For SBTs and simulation-enabled performance tasks, potential unfairness due to Task $\times$ Person interaction may also be present (Linn & Burton, 1994; Shavelson et al., 1993).

Task $\times$ Person interaction occurs when one test taker is more familiar with, knowledgeable about, or interested in a particular context or topic than another test taker. For shorter tasks, this effect can be easily balanced by having a variety of different contexts or topics across items. SBTs and simulation-enabled performance tasks, however, tend to take significant time, making it impractical to include more than a few tasks. For example, the NAEP Technology and Engineering Literacy assessment tasks take 10–30 minutes each, leaving little room for additional items with contextual or topical variations (a problem NAEP addresses through its matrix design and the reporting of results at the group level, rather than through individual scores).

For TEIs and many extended-interaction items, well-understood psychometric models such as item response theory (IRT) can be used because student response data can be reasonably assumed to meet the statistical assumptions of such models. Any media or interactivity, however, must be carefully evaluated to ensure that their addition does not introduce extraneous cognitive load or other forms of construct-irrelevant variance. Particularly for extended-interaction items, the ease with which the computer interface is learned and employed can significantly affect performance. To ensure fairness, assessment developers must consider the range of experience levels likely in the target population. Similarly, items that use video or audio, or that have fine-motor-skill requirements, must be created to be accessible from the outset. Universal design methodologies are the de facto standard for achieving this goal, including such features as braille, stacked multilingual translations, videos in American Sign Language, and glossaries and test directions in other languages (e.g., Smarter Balanced, 2024).

SBTs and simulation-enabled performance tasks raise significant issues for modeling and statistical inference. While SBTs often contain a set of scored items that resembles a traditional assessment, performance on these items may not necessarily be statistically independent conditional on proficiency. The overarching narrative and topic that define the SBT contribute one source of local dependence that may affect estimates of test precision under standard analysis approaches. When multiple SBTs are used within a larger assessment, a testlet model (Wainer et al., 2000) can be employed to account for the added dependencies among item responses within an SBT.

Another source of dependence might come from how items are related over the scenario. Test takers who draw the wrong conclusions in early items may carry those incorrect ideas forward. Such is particularly the case when the SBT is structured so that later items build on previous results, as in the bike lane task. The effect of previous errors can be mitigated through leveling, in which the test taker is given a correct starting value for the new item. Leveling is controversial, however, because it does not allow test takers to follow their own path, nor does it necessarily eliminate the dependency that could be caused by the test taker's memory of previous reasoning. Models that can

accommodate item dependency, such as Bayesian networks, might be suitable in this context (Almond et al., 2015).

Simulation-enabled performance tasks often have a significant number of correct responses, especially when the response process itself is key to the target construct. In such cases, the log files contain fine-grained details such as mouse clicks, keystrokes, and latencies, which on their own are likely to have very limited meaning. Lending meaning requires thoughtful planning and development, starting with task design (Clauser et al., 2016). Such an approach involves creating a task so that it elicits important features of performance, which can then be aggregated in some fashion for scoring. This approach was used in the USMLE and in the design section of the Architect Registration Examination (Bejar, 1991; Braun et al., 2006; Clauser et al., 2016). New analytic techniques, such as deep learning (LeCun et al., 2015), or novel applications of cognitive-process modeling (LaMar, 2018) may also help in meeting the scoring challenge. The use of response-process data will be discussed further in the section “Response Processes” and automated scoring will be discussed in the section “Automatically Scoring Complex Constructed-Response Tasks.”

Comparability of Results and Score Meaning

As noted earlier, many current TBAs can be characterized as first or early second generation—that is, they differ modestly from their paper-based predecessor or current counterparts with respect to item types and the competencies targeted for measurement. However, some of these TBAs also make use of innovative items. Thus, they are beginning to leverage the affordances associated with digital administration. In this section, we deal with the implications of such leveraging for the comparability (i.e., continuity of meaning) of results.

In one use case—national assessment—the comparability of results from new TBAs to prior paper-and-pencil (PBA) results is generally desired. Comparability to PBA versions is sought because one of the defining characteristics of NAEP is the ability to compare the performance of students in the early 21st century—overall and for various subgroups—to cohorts from prior decades. In another common use case, K–12 accountability as implemented through Smarter Balanced and PARCC, the TBAs replace prior, distantly related, paper tests with no expectation of comparability. However, each of these TBAs may coexist with its own PBA counterpart, the choice of examination mode being left primarily to states, districts, and schools.

In both use cases, test takers and test users alike may desire, or expect, that results obtained under either examination mode should be comparable to the other. For group-score assessments like NAEP, recent results obtained with TBA—for the United States as a whole, for demographic subgroups, and for states and districts—need to provide meaningful comparisons to the results from earlier years in which the administration mode was paper and pencil. Changes in scores need to be interpretable because of changes in the target competencies (i.e., to construct-relevant factors), not because of an artifact of the switch from paper to digital delivery. Similarly, for K–12 accountability

tests, differences in results between districts, schools, classrooms, or individual students should reflect differences in the constructs of interest, not factors related to examination mode.

Moreover, in some use cases—for example, TBAs for K–12 accountability—the TBAs themselves allow for considerable variation in administrative conditions. Both the Smarter Balanced and the PARCC tests, for example, are delivered on a range of laptops and tablets, and student responses can be input through keyboards native to the device, an external keyboard, or, in some cases, touchscreens. Despite this variability, test users and takers again expect the results—individual and aggregate—to be comparable across these variations.

The definition of what it means for scores from different versions of a test, or from different tests with similar targeted competencies, to be comparable has received much attention in the measurement literature, initially within the context of more general discussions of score linking and equating (Holland, 2007; Holland & Dorans, 2006; Linn, 1993; Mislevy, 1992; see also Moses, this volume). In this literature, comparability is viewed as a matter of degree, with the strongest level being achieved when scores from different tests can be considered *interchangeable*. Less-stringent degrees are achieved through other forms of scale alignment such as concordances, statistical relationships between the results of assessments of related constructs that hold in specific populations.

Interchangeable scores can often be obtained in the context of consequential tests, like SAT and ACT, which traditionally produced different parallel forms administered in the same mode (primarily PBA) for use at each scheduled test administration. Such forms are constructed so that the assessed content and psychometric characteristics are tightly controlled to enhance comparability. As a result, the test takers at a given level of proficiency can be expected, on average, to achieve the same test score and to be measured with the same degree of precision, regardless of the test form administered. Moreover, score meaning—in terms of the competencies measured—can also be assumed to be the same, regardless of the form given.

It is noteworthy that even within this highly constrained context, it is rarely the case that the *raw scores* from different forms (e.g., simple number-correct scores or even IRT-based estimates) can be treated as interchangeable. However, with proper data collection designs and appropriate analysis procedures, scores from these alternate forms can be *equated* (i.e., adjusted for any unintended differences in difficulty and/or expressed on form-invariant scales, like the well-known ACT and SAT scales). For most practical purposes, after equating, the resulting scale scores from different forms of the same test can be treated as interchangeable in terms of the constructs measured, as well as psychometric characteristics, when making inferences about individual test takers or groups.

In considering the comparability of scores from PBA and TBA forms, interchangeability may be a reasonable and desirable goal when the content, item types, and delivery paradigm (e.g., linear versus adaptive) of the two modes are kept as close as possible. In situations

where the identical (or nearly equivalent) set of items is being delivered, there is a long history and considerable body of empirical evidence with different kinds of tests and testing populations, often referred to collectively as *mode comparability studies*. Though findings from this body of research are somewhat mixed, the evidence suggests that—as is often the case with different PBA forms of the same test—some degree of equating adjustment to TBA raw scores may be required if results are to be treated interchangeably with those of the corresponding PBAs. Papers by Paek (2005), Kingston (2009), Drasgow et al. (2006), Lottridge et al. (2010), and W. D. Way et al. (2006) offered pertinent reviews for tests that provide individual scores. Jewsbury et al. (2020) gave an additional recent example from NAEP pertinent to group-score assessments.

In some use cases (e.g., the Smarter Balanced K–12 accountability tests and the PISA transition to TBA), design differences between PBA and TBA also exist—for example, the paper version being linear and the computer one item-level or multistage adaptive. Changes in test delivery that involve an item-level CAT or an MST could, in some cases, be expected to affect the psychometric properties of scores, usually by improving the precision of results for low- and high-performing test takers relative to test takers in the middle of the score distribution. Hence, strictly speaking, scores from the linear PBA and adaptive TBA may not be interchangeable in the same sense as scores on parallel forms from the PBA test, the differences in precision at particular ability levels being intentional and desirable. However, through the imposition of content and format constraints and the conduct of appropriately designed equating studies (see, e.g., Dorans, 2000; Schaeffer et al., 1995, 1998), adaptive versions of linear PBAs can be instantiated such that scores can be treated as equivalent measures of the same construct for practical purposes.

As of this writing, however, the TBA versions of a test can be expected to differ at least to some degree from their PBA counterparts in ways that go beyond a simple move to adaptive delivery. Even first- and second-generation TBAs frequently contain some number of TEIs for which exact PBA counterparts do not exist. As a result, subtle cross-mode differences in the assessed target competencies are inevitable. Similarly, construct-irrelevant factors—such as familiarity with and ease of working within the digital and paper testing environments, respectively—might also reduce the degree of comparability. Different approaches to scoring constructed-response items represent an additional threat to interchangeable scores. For example, absent training, human raters have been found to differentially grade handwritten versus computer-entered responses (Russell & Tao, 2004a, 2004b; Sandene et al. 2005, pp. 14–15).

The *Standards* (AERA et al., 2014, p. 105) requires a clear rationale and supporting evidence for claims that scale scores earned on alternate forms of a test may be used interchangeably. Given the goal of leveraging the affordances of digital delivery, achieving interchangeability is unlikely, perhaps unnecessary in many use cases, and probably undesirable if it limits improvements in measurement precision and better representation of important constructs. However, even when strict equivalence is not claimed, the *Standards* requires a rationale and direct evidence of the *degree* of

score comparability commensurate with intended uses and claims (p. 106). Similarly, U.S. Department of Education guidance on peer review of state assessment systems (U.S. Department of Education, 2018) suggests that when a state administers different versions of its assessment (e.g., TBA and PBA, or TBA on different devices), the state should provide comparability evidence generally consistent with the expectations of current professional standards.

What sorts of evidence traditionally have been, and should be, provided to document the degree of comparability between TBA and PBA results? Whereas the *Standards* contains no specific guidance, there is a substantial literature that addresses the topic generally. Some of this literature is located within the broader context of establishing comparability of scores obtained under standardized conditions to those obtained from test variations such as accommodations, adaptations, or different languages, where interchangeable scores are rarely practically achievable. Particularly pertinent discussions can be found in Kolen (1999), Wang and Kolen (2001), Sireci (2005), Winter (2010), Lottridge et al. (2010), Randall et al. (2012), DePascale et al. (2016), and Berman et al. (2020).

There is considerable agreement in this literature regarding the sources of evidence required. These sources include studies that examine:

- similarity of the dimensional structures;
- similarity of item-level psychometric properties (e.g., classical test theory indices of difficulty and discrimination, IRT curves, IRT model fit in cases where the TBA and PBA versions contain identical or corresponding items);
- similarity of predictive and concurrent statistical relationships between assessment scores and other related educational variables;
- similarity of measurement precision through comparisons of overall standard errors of measurement and standard error curves across the proficiency range;
- similarity of overall score distributions, including means, degree of dispersion, and shape; and
- similarity of score differences and extent of differential item functioning for important subgroups defined by gender, race/ethnicity, disability, socioeconomic status, or computer familiarity.

Whereas a comprehensive evaluation of comparability would involve all these sources of evidence listed, rarely in practice is that possible. In analogous fashion with validity, degree of comparability remains an integrative judgment by test professionals, test users, and other stakeholders as to whether claims based on comparisons of PBA and TBA versions are adequately supported.

From a fairness perspective, perhaps the most important sources of evidence for tests used for consequential purposes are associated with what Winter (2010) labeled “score-level comparability” (the last three sources listed above). Direct evidence is usually gathered from mode comparability studies in which TBA and PBA results are obtained from groups that can be assumed to be equivalent with respect to the

distribution of the target competencies associated with the test. Data from such studies serve two roles. The first role is to provide evidence of score-level comparability. The second role is to, when necessary, determine an adjustment from one or the other modes to account for the effects—that is, differences in overall test performance (means, standard deviations, or shapes). Depending on other sources of evidence, the adjustment can, in the best case, render the scores effectively interchangeable or, where claims of interchangeability are not justified, increase the degree of comparability and fairness inherent in score use.

Mode-Study Data Collection Designs

Mode comparability studies have made extensive use of the data collection designs traditionally employed for equating and linking (see, for example, Kolen & Brennan, 2004; and Moses, this volume). *Single-group designs* have many advantages, in principle. In these designs, test takers are given both the PBA and the TBA versions, usually with the order of administration randomly counterbalanced (e.g., Gallagher et al., 2002). Because the same test takers are tested in both modes, issues related to sampling variability between groups are minimized. Moreover, this design provides the most direct evidence of the construct equivalence of the TBA and PBA versions. The correlation between scores in the two modes and the similarity of internal dimensional structure, psychometric properties, and relationships with other variables can all be directly evaluated *on the same group of test takers*. If IRT methods are employed, the appropriateness of jointly scaling both tests can also be directly appraised by examining model fit.

However, there are several challenges with implementing such designs. It is often not feasible to administer tests in both modes to the same test takers because of concerns about burden and so-called *order effects* (i.e., the fact that the relationships between the TBA and PBA versions may differ depending on administration sequence). Lottridge et al. (2010) described several such studies where differential order effects were found.

A somewhat more practical alternative that offers many of the advantages of single-group designs is the *random-groups design* (i.e., where test takers are randomly assigned to administration mode). Because test takers are assessed in only one mode, issues of burden and order are not relevant. When data from large, randomly equivalent groups of test takers are available, almost all the sources of evidence recommended above can be produced, the major exception being correlations and joint IRT scaling. NAEP has relied, and continues to rely, on such designs in carrying out its transition from PBA to TBA (e.g., Bennett et al., 2008; Jewsbury et al., 2020). As part of its 2015 field test, PISA also employed this design to aid in its TBA transition (von Davier et al., 2019).

Operational research from NAEP, described in detail by Jewsbury et al. (2020), is a particularly good example. Data from the random-groups design were used to adjust for mode effects associated with TBA, as well as to provide comprehensive analyses (including classical test theory and IRT), documenting the evidence for the comparability of NAEP results across modes after adjustment. In this study, NAEP administered

the PBA and TBA versions of its fourth- and eighth-grade 2017 mathematics and reading assessments to random samples of test takers, with random assignment to mode carried out largely within each NAEP-participating school. This *within-school* design was chosen to produce highly similar samples, taking each mode at all the jurisdiction levels for which NAEP reports results (nationally, by state, and for large urban districts), and to give reasonable statistical power for detecting mode differences at the state and subgroup levels.

National-level results comparing item-level statistics across modes (biserial correlations, differential item functioning statistics, IRT parameter estimates), as well as various other psychometric characteristics (e.g., dimensionality), provided solid evidence that the two versions were measuring highly similar target competencies. However, other test-level analyses indicated that, without adjustment, results from the TBA versions of the 2017 reading and mathematics assessments would be systematically lower for both fourth- and eighth-grade test takers, with larger mode effects for fourth graders. After making a *single adjustment* to the TBA NAEP scale-score distributions to equate the mean and standard deviation for the national samples taking the assessment in each mode, results were shown to be highly similar for the major national reporting subgroups (e.g., sex, race/ethnicity), for states, and for participating large urban districts. That is, with very few exceptions, the observed differences due to mode were within the bounds of sampling error. The few statistically significant differences that were observed showed little consistency across grades and subject.

PISA employed a somewhat different approach (von Davier et al., 2019). In the 2015 field test, a number of countries provided data from randomly equivalent samples of 15-year-olds for both PBAs and TBAs of reading, mathematics, and science. The investigators combined descriptive analyses—that is, visual inspection of residual plots and indices of item-level model fit—with statistical evaluation of a series of constrained IRT models that they refer to as “mode-effect models.” The mode-effect models imposed increasing degrees of measurement invariance. These analyses identified some items for which comparable functioning could not be supported. However, for most items, the evidence suggested that the IRT parameters were equivalent across modes. These items were then treated as equivalent in the operational 2015 analyses so that results from both PBA and TBA could be reported on a common scale comparable to prior PISA assessment cycles (OECD, 2017, chap. 9).

Whereas the random-groups approach can be quite effective, it too may frequently be intractable because it relies heavily on random assignment and requires large sample sizes for sufficient statistical power. Whereas NAEP was able to achieve both requirements, those requirements may not be easily met in other operational contexts (e.g., where the test-taking population has uneven access to computers or where imposing an administration mode is not acceptable to the test taker or institution). In such situations, data collection by necessity will involve self-selected, *nonequivalent groups*, that is, groups in which similarity of the distribution of target competencies for the TBA and PBA versions cannot be assumed.

With nonequivalent groups, mode comparability studies have typically adopted one of two strategies (or their combination). The first strategy is to use statistical approaches that employ demographic and other related performance data to control for preexisting group differences in the targeted constructs. Approaches such as propensity-score matching (Rosenbaum & Rubin, 1985), coarsened exact matching (Iacus et al., 2011), or weighting-based approaches (Haberman, 2015) can be employed to create *pseudo-equivalent groups*. These matched groups can then be used to conduct the desired comparability analyses, including the calculation of any mode adjustment. Numerous examples of this approach can be found (Lottridge et al., 2010; W. D. Way et al., 2006, 2007, 2008; Yu et al., 2004).

A second strategy is to rely on *common-item assumptions*—that is, the assumption that identical or nearly identical items appearing in both modes exhibit the same psychometric characteristics (e.g., difficulty and discrimination). When employed in the context of nonequivalent groups, this strategy can be thought of as a variant of the *common-item* or *anchor test nonequivalent groups design* (Kolen, 2007). Numerous IRT methods can be applied, including concurrent calibration and approaches based on separate calibration of PBA and TBA data, followed by some form of parameter linking on test characteristic curve transformations. In addition, non-IRT methods can be brought to bear (see, for example, Kolen & Brennan, 2006, chaps. 4–6). Usually, the analyses proceed by first assuming that all identical/similar items function comparably and then selectively relaxing this assumption based on various diagnostic indices, such as model fit and analysis of outliers.

As a basis for evaluating comparability, both approaches to the nonequivalent groups situation come with challenges that can compromise their effectiveness. Statistical matching requires sufficiently strong ancillary data related to test performance that accounts for the differences between the groups with respect to the targeted competencies. When such ancillary data are not available, the matching will not produce equivalent groups. As such, conclusions regarding comparability, or adjustments to achieve comparability based on equating the score distributions in the matched groups, may be suspect. It is important to note that ineffective matching could potentially affect conclusions in both directions. That is, one could be led to conclude that scores are not comparable when in fact they are or to miss a systematic directional bias that was inadvertently removed through faulty matching.

A possible example of this situation can be seen with the PARCC K–12 reading and mathematics accountability tests, which existed in both TBA and PBA form. Though not strictly identical with respect to all item types, the two modes were intended to produce comparable results. Comparability studies were conducted based on field test data in 2014 (Brown et al., 2015) and on the first year of operational data in 2015 (Liu et al., 2016). The first of these studies was originally intended to use a random-groups design, but challenges in implementation made it necessary to rely on a post hoc matching approach. In contrast, the second study was designed to employ nonequivalent groups with propensity-score matching. In each study, a comprehensive analysis was done

using most of the generally recommended evidence types (AERA et al., 2014). Both studies presented solid evidence that highly similar constructs were being assessed by the two modes. Moreover, though small mean differences usually favoring PBA were noted for particular test subject/grade combinations, the differences in average scores between modes were, for the most part, negligible. Thus, the general conclusion was that results could be compared across modes for accountability purposes.

Despite this conclusion, subsequent studies and analyses of operational PARCC results have suggested that, in at least some districts and states, scores on the PBA version were systematically higher to a greater degree than would have been expected from the mode comparability study results (Backes & Cowen, 2018; Duque, 2017). One reason that evidence of noncomparability has surfaced may be that the earlier matching procedures did not work as intended, a possibility expressed in the 2016 PARCC Technical Report (Pearson, 2017, pp. 143–144). However, other considerations—such as differences in the relationship between TBA and PBA results across groups, states, and time—represent alternative, or at least contributory, factors. Several of these considerations are discussed in the following paragraphs.

Relying on common-item assumptions effectively assumes that, at least on average across all the items common to both modes of presentation, consistent *main effects* on item difficulty and discrimination favoring one or the other mode are negligible. Studies employing this approach are most typically carried out within an IRT framework. In one variation, IRT item parameter estimates based on data from one mode (the reference mode) are applied to the data from the other mode. Analyses of model fit are done to identify items for which the assumption of common parameters appears untenable, and these items are then deleted from the set. For these identified items, as well as items unique to the other mode, separate item parameter estimates based on data from the new mode are obtained, with the parameters for the remainder of the items fixed at their reference-mode values. The full set of item parameters is then used to produce results that are directly on the reference-mode scale.

In situations where the tests have been carefully created to measure similar constructs across modes and the assumptions of negligible main effect are tenable, this approach can achieve comparability of score distributions, as well as provide clues regarding item features that interact to introduce cross-mode noncomparability (e.g., when the mechanics of response entry for a TEI are more complicated than answering the analogous question on paper). The latter information can be quite valuable in producing future comparable versions. However, the assumption of negligible main effects cannot be effectively evaluated based on the mode-study data collected with the non-equivalent groups design. This situation stems from the fact that main effects on item parameters due to mode are not separable in the IRT analysis from group differences on the target construct since the two causes are perfectly confounded. Moving forward, these differences will manifest themselves as a systematic source of noncomparability between scores in the two modes. Attempts are sometimes made to combine the common-item approach with matching strategies to disentangle the mode effects from

group differences (ETS et al., 2016; Liu et al., 2016). However, the effectiveness of this strategy is subject to the limitations discussed previously.

Much of the conceptual framework and methodologies within which psychometricians have approached issues of comparability (particularly of score distributions) was developed when consequential testing was done as PBA under generally strict standardization of administration conditions and testing formats. To be sure, testing conditions across PBA sessions varied in small, presumably inconsequential ways. The basic PBA administration infrastructure could be assumed to be stable, reasonably homogeneous, and familiar to the vast majority of current and future test takers. These assumptions of homogeneity and stability undergirded the argument that standardized test scores were fair, valid, and reliable and that results from such assessments could be confidently and meaningfully compared over time.

Early forays into consequential TBA (e.g., for military selection, licensure, graduate admissions) occurred largely with adults at test centers where the variability in equipment (desktop computers) was limited and under the control of the testing organizations. Given these conditions, it was also reasonable to assume that the TBA administration infrastructure was stable and homogeneous. Thus, similar claims for the fairness, validity, and comparability of TBA results could be made. Against this backdrop, historical approaches to evaluating comparability (particularly of score distributions) seemed appropriate. Unintended differences in difficulty due to mode, for example, could be identified and adjusted for *once*, much the way adjustments have been made to alternate forms of admissions tests like the SAT and ACT. That adjustment could then be applied to results produced by future forms of the TBA, rendering them comparable to past PBA results and to one another.

However, the technology landscape for TBA is different and far more variable, particularly in K–12, where laptop computers and tablets are used for instruction as well as testing. Differences in displays (size, resolution), input mechanisms (keyboards, mouse, touchscreen), and operating systems carry the potential to inadvertently introduce construct-irrelevant differences and impact psychometric properties, not only in comparison to PBA versions, but also among the “same” TBA taken on different devices.

NAEP provides a good example of the challenges that consequential testing programs face. To remain relevant, NAEP has already transitioned many of its assessments (reading, mathematics, U.S. history, civics, geography, and science). Up through 2024, NAEP’s operational approach has been to standardize by administering the assessment on the *same* device configuration brought into the participating schools by NAEP personnel. This approach is an attempt to provide, at any given time, a stable and consistent infrastructure capable of delivering the full range of NAEP assessment tasks and of producing comparable scores.

However, maintaining a stable delivery architecture *over time* is becoming untenable. Hardware, software, and interface design life cycles make change inevitable. New delivery devices, operating systems, and assessment software are periodically

introduced. Concomitantly, as NAEP gains experience with its digital delivery system, the program itself will want to make user interface and task design improvements. These changes will affect to varying degrees how the test taker interacts with the test, which could in turn materially change the meaning of results over time. This evolution will be continuous and inexorable.

Of course, when NAEP was a paper-and-pencil assessment, changes in assessment procedures and content were also periodically introduced. At these limited change points, *bridge studies* (using random-groups designs) were conducted to adjust for the potential impact of the changes on assessment results and to confirm the valid reporting of trend. For PBA administrations, results from the sequence of assessments occurring before and after the change could be safely assumed to be comparable since the basic underlying test delivery system (e.g., booklets, pencils) represented a constant delivery infrastructure. However, in the TBA era, conducting mode-comparability studies at major change points may not be practical, economically feasible, or enough to confirm valid trend reporting. Even with a relatively stable assessment in content and procedures, the constant evolution of the delivery infrastructure may be too great.

In contrast to NAEP, Smarter Balanced and PARCC have taken a different approach, electing to support TBA delivery across a wide range of digital devices. This strategy presents a different challenge: how to accumulate and present evidence of comparability across a potentially large number of different devices. As is true for mode effects, the research on device effects has generally used some version of the designs described earlier. Reviews by DePascale et al. (2016) and W. D. Way et al. (2016) summarized some of the key findings about the impact of such factors as screen size, input device, and item type, as well as how these factors interact with such things as content area. But given the continuing rapid evolution of TBA, a strategy of amassing comparability evidence (as historically done for PBA and TBA) would appear to be Sisyphean considering the plethora of existing TBA device configurations. And as noted, the COVID-19 pandemic brought the administration of some TBA programs to the home, where an even greater array of device configurations may be found.

Considering this reality, a more expansive approach to comparability is required. W. D. Way et al. (2016) argued that we have moved from an era of standardization to one of personalization. In that latter era, consequential testing is more accurately viewed as a collection of variations, many of which are intended to adapt the experience to the individual in a way that maximizes their access to assessment content (see also Bennett, 1999, regarding “generalized accommodation,” and Bennett, 2024). From this point of view, it will be necessary to supplement, if not entirely replace, studies documenting the comparability of results between variations with mixed methods interdisciplinary approaches. Such approaches should provide test, item, and interface design principles usable across a wide range of assessment situations. The knowledge obtained would inform the engineering of personalized testing conditions in a way that maximizes validity and fairness by increasing the likelihood that results can be treated as comparable across a range of devices.

As W. D. Way et al. (2016) so aptly put it,

What is clear is that the evolution of technology will only continue and that schools will continue to adopt different technologies in different time frames and will have little patience for a measurement field that is unprepared to accept these technologies for testing purposes. Establishing a framework and a process for evaluating new devices and new technologies is perhaps more important than understanding the impact to comparability of any specific device or technology. (p. 274)

D. Way and Strain-Seymour (2021) have taken a step in that direction by proposing a framework for device and interface features that might affect test performance and, thereby, perhaps offer a path to greater comparability through the personalization in the manner W. D. Way et al. (2016) suggested. That idea is consistent with an emerging alternative view of standardization in educational testing more generally. For example, Sireci (2020) argued for UNDERSTANDization, in which the first step is to appreciate the implications of diversity within the student population and then critically evaluate the ways in which traditional standardized procedures may lead to biased estimates for some individuals and groups. The third step involves adjusting those procedures to eliminate potential biases.

Some precedent for this view exists in the way in which testing programs now handle accessibility. For example, Smarter Balanced offers accessibility features in three categories: accommodations (available to those students with documented need through either an Individualized Education Program or a 504-accommodation plan), designated supports (available to any student for whom school officials have indicated the need), and universal tools (available to all students; Smarter Balanced, 2023). Universal tools include English glossary (i.e., pop-up definitions for selected construct-irrelevant words), highlighter, strikethrough, zoom, and notepad. Among the designated supports are masking, color contrast, text to speech (for all items except reading passages), glossary translation in 10 languages and several dialects, and translated test directions in 19 languages. Accommodations are provided through braille, closed captioning (for listening items), American Sign Language video presentation (for listening and math items), and text to speech (for reading passages in Grades 6 and above), among other mechanisms.

Smarter Balanced test administrations, thus, may vary considerably across students depending on the category of accessibility features for which they are eligible and the features utilized within that category. This variation is intended to follow the guiding principle that, when selectively metered, accessibility supports can contribute to more valid measurement because the assessment is more appropriately customized to student characteristics.

The beginnings of a scientific framework within which to conceptualize comparability for students with special needs may exist in the research on the so-called interaction hypothesis and on differential boost. The interaction hypothesis indicates that accommodated tests (e.g., the availability of text to speech) should result in students

with special needs achieving higher scores than they otherwise would, coupled with no difference for general education students taking the same accommodated test. In contrast, differential boost focuses on the expected increase in scores for students legitimately needing the accommodation, ignoring any effect (or lack thereof) on general education students. Sireci and O’Riordan (2020) provided a comprehensive review of issues related to accommodations and comparability, Sireci et al. (2005) gave a discussion of the interaction hypothesis among students with special needs, and Pennock-Roman and Rivera (2011) did the same for English learners.

Though comparability is emphasized in many consequential use cases, it is important to acknowledge that too large an emphasis may have the unintended negative consequence of inhibiting advances in measurement science and practice. Comparability of scores to previous versions of an assessment may be undesirable if it is obtained at the cost of innovation in test design and task types that provide the potential for improved measurement precision and tapping important constructs. Thus, as testing programs transition from PBA to TBA, they must carefully weigh the costs and benefits of maintaining comparability to past results.

When severing comparability to past results, new reporting scales are typically introduced, which can be extremely disruptive to test takers, test users, and other stakeholders. Testing programs that do so often provide concordance tables showing pairs of scores, new and old, having the same percentile rank in a particular population of test takers. Such statistical relationships can be helpful during the transition when stakeholders may be required to make consequential decisions about test takers, some of whom have scores on the prior reporting scale and others on the new scale. Concordances, however, carry their own risks of misuse and misinterpretation. The reported statistical relationship does not imply interchangeability of results and score meaning. In situations where the knowledge, skills, and abilities measured by the old and new version of the test are substantially different, the statistical relationships may not hold in populations different from those on which the concordance was established or in subgroups of the population. (See, for example, Dorans & Walker, 2007; Pommerich, 2007; and Sawyer, 2007, for a discussion of the uses and limitations of concordances.)

Strategies like instituting innovations incrementally to avoid disrupting comparability at a single time point represent a different approach to balancing the trade-offs. The strategy being used by NAEP for its assessments in reading, mathematics, and science appears to follow such an evolutionary approach. In its initial stages, the transition concentrated on moving the current paper assessment to digital delivery. Later stages introduced innovative item types like SBTs and simulation-enhanced performance tasks. Subsequent stages include a 2025 field test of web-based delivery on school-provided devices and possibly further explorations of adaptive testing.

Introducing such innovations in stages—and evaluating their impact on the comparability of results through careful experimental study and other empirical approaches (see, for example, Jewsbury et al., 2020)—will help NAEP develop validity evidence for supporting construct-based inferences from existing trend lines. If supported

by empirical study, changes in results from the original assessments to the newer innovative ones can be interpreted as due largely to construct-relevant factors, overall and for most or all subgroups of interest. The evidence supporting construct-based inferences is likely most sound for those assessments close to the transition points—and to the comparability studies—where the innovations were introduced. The validity of construct-based inferences may be less well supported for assessments farther apart in time, given the potential cumulative effect on score meaning of the multiple innovations and the absence of studies directly comparing the modified and original versions. Given the rapid changes in technology and the potential advances in assessment practice, increased ambiguity as to the meaning of changes in results over longer time spans may be the price paid for maintaining a relevant assessment system.

In other instances, where an evolutionary approach is not practical or desirable, breaking with the past may represent the wiser course. NAEP again provides an example where decisions to forego comparability with prior assessments were made. The most recent instance concerns the transition of the writing assessments at Grades 8 and 12 from paper to TBA in 2011 (U.S. Department of Education, 2012). The new NAEP writing assessment was based on a different framework than that used in prior years (1998, 2002, and 2007). The implementation of that framework required the assessment to reflect the reality that most writing in the second decade of the 21st century in educational and work settings was already done with the aid of technology. TBA delivery included the concomitant introduction of resources such as a thesaurus, as well as such common computer tools as spell check, cut, copy, and paste—all of which had no analogue in the NAEP PBA context. In addition, the new framework called for changes to the assessment tasks that required a purpose for writing and a specific audience to be addressed. NAEP decided that the changes reflected important social and educational developments and chose to begin new trend lines without empirical study of comparability. The loss of comparability to prior assessment results was seen as a necessary and desirable trade-off in return for maintaining the program's relevance and position as an innovation leader.

Response Processes

Whereas a loss or reduction in the comparability of results and score meaning to PBA might be viewed as a negative effect of the transition to TBA, a more positive outcome relates to the availability of data about the response process. The response process describes the cognition in which a test taker engages when encountering an item stimulus, formulating an answer, and entering that response (Wilson, 2005; Ercikan & Solano-Flores, this volume). For some constructs, the test taker's cognitive process is itself a relevant target of measurement, either instead of, or in addition to, the final product of that process (e.g., in medical patient management, scientific inquiry). This section discusses how technology enables the observation and analysis of evidence concerning response processes. Such evidence can be captured and analyzed from the data collected in any type of TBA, even ones that include only traditional

multiple-choice items. However, the potential value, as well as the challenges of such capture and analysis, increases with task complexity.

The simplest response processes might involve multiple-choice tests of declarative memory: The respondent is asked to recall a fact, find the matching response from a fixed list, and mark the appropriate option. The response process to an algebra item might be more complex, perhaps involving multiple steps of problem translation, integration, solution planning, and execution. A yet more complex response process is evoked when a student is asked to write a persuasive essay. This task might bring into play such aspects as choosing the thesis, outlining and planning the argument, drafting text, and editing and revising that text. A final type of response process is exemplified by a simulation-enabled science performance task. Here, selecting a hypothesis, choosing variables, designing and running an experiment, and interpreting the results would all be relevant.

Next we describe the diverse uses of response process data, after a short introduction to the types of data that TBAs can collect. These uses include validation, quality control, security, and new insights into group and individual performance.

Types of Response Process Data

TBAs can, in principle, provide a stream of process data to supplement the traditional scored response. Log files are commonly generated, recording key events in the response, as well as in the navigation of the assessment. Design of good log files is a new but important part of assessment development (Hao & Mislevy, 2018). On the one hand, whereas every keystroke, finger swipe, mouse action, and latency can be recorded, the resulting data will constitute a large, potentially uninterpretable, collection. Recording too little, on the other hand, may lose valuable information. In general, it is recommended to record as much information as possible because new analytic techniques are emerging for dealing with the data.

No matter the actual item format, TBAs can provide latency data for each student on each item. For example, when a student loads an item, when they select a response, and when they submit can be recorded. For reading comprehension items, timestamps can denote how long a test taker stayed on a single screen; for a writing task they can indicate how long the test taker paused before beginning to write and between characters, words, and sentences. The more interactive an item, the more timestamped events can be collected.

As the called-for interaction increases, more details of the process used to complete a task can be inferred from the log file. In writing tasks, for example, the recorded keystrokes can allow for a real-time recreation of exactly how an essay was composed. Mouse actions can also be recorded for items that involve more complex response formats. For highly interactive tasks, such as simulation-enabled performances, each step can be logged, providing a fine-grained record of the test taker's path to solution.

It is useful to consider the different levels of inference that might be made about a test taker's response process from log files. The lowest level is descriptive, based on the raw

data. These data might include each keystroke, cursor movement, and mouse action, along with assessment-generated events such as the running of a video clip. While data at this level are often more difficult to interpret, they may be used to create a playback, or description, of a student performance. Such playbacks can be used for quality assurance, validity studies, and performance ratings by humans and as a formative resource for instructors or test takers. This level of data is also objective. If the log records a mouse click on a specific button at a particular time, it is almost certain that the click occurred as recorded.

A next level of inference is the action or decision layer. At this more semantic level, mouse clicks or keystrokes may be interpreted as indications that the test taker decided to take an action within the task. Thus, a mouse click becomes “selecting response X” or, in the context of a science simulation, “running a new trial.” While these inferences are relatively low level, one cannot be sure that the test taker intended the interaction as it might be interpreted.

The highest semantic level identifies strategies, plans, and knowledge. This level requires the combination of multiple events into a meaningful pattern in the context of the task. For example, we might infer that a student implemented a “control-of-variables” strategy to test a particular hypothesis (LaMar et al., 2017) or that a student wrote an outline as a plan for their intended essay.

It is worth noting that timing data can strengthen or weaken inferences. For example, we might infer that a test taker did not take the time required to develop a strategy or even read the question prompt. This judgment would indicate that the events recorded in the log file were more likely associated with random noise than with planful actions and strategies.

Utility of Response Process Data

Because the test-taker response process is central to the integrity of educational measurement, process data have multiple uses throughout the assessment life cycle. Assessment designers have long employed student cognitive labs to test and refine items in development (e.g., see Connolly & Wantman, 1964, for an early example). Think-aloud protocols or retrospective interviews aim to make observable how the test taker works through questions so that item functioning can be improved.

With the help of automatically generated process data, far larger test-taker samples can be evaluated than are possible with these more labor-intensive methods. In the early stages of assessment design, such data can be used to identify usability problems with the computer interface and to select which tasks are most likely to evoke response processes involving the target construct. Items that allow for shortcut solutions or the use of construct-irrelevant skills can be identified and modified or dropped. Whereas gathering some types of validity evidence is a routine part of assessment design (e.g., item alignment, cognitive labs), most evidence is generally gathered from pilot, field test, and operational administrations because those events provide the sample sizes needed for psychometric analysis. Examining automatically gathered process data could

complement these analyses (Ercikan & Pellegrino, 2017). For example, Zhang, Deane, et al. (2019) compared the psychometric properties of a scenario-based English language arts assessment design to alternative structures and found that a single scenario functioned best with the essay appearing after (rather than before) a series of lead-in tasks. These psychometric analyses were complemented by process investigations that evaluated *how* students wrote their essays. Those investigations showed that the scenario-based design reduced the impact of general writing fluency processes on essay score, thereby presumably giving students with less keyboard facility or lower verbal fluency more opportunity to display their argumentation skills (Guo et al., 2020; Zhang et al., 2017). (See Padilla & Benitez, 2014, for a detailed discussion of theory, relationship to other evidence types, and methods in using processes for validation.)

Once an assessment is deployed, response process data can be a valuable tool for test security and data integrity (Qian et al., 2016). For example, item response times can be used to identify unusual patterns that may suggest cheating (Marianti et al., 2014). Test takers who use a hidden answer key or memorize answers from an illicit examination copy are likely to have a mismatch between their item completion times and the distribution of times from the test-taker population. Such a mismatch may occur in part because, by virtue of not having to engage item solution processes, the former group moves through items more quickly than do honest test takers. Similarly, a test taker who is copying from another test taker will show not only great similarity to that individual's responses, but also synchronicity in timing with those responses. The detection of any such events can allow the possibility of real-time alerts to examination proctors.

Process data can also be used as part of quality assurance to identify when something has gone wrong in administration or scoring. With the addition of technology, quality assurance must include digital delivery, user-interface functionality, data recording, automated scoring, and data transfer and storage. Increasingly, assessment developers look to the best practices of software engineering for proper quality assurance methodologies because the assessment is a software product. Similar to how software engineers create unit tests during development, assessment developers can specify constraints on the expected response patterns while they craft the items. Cognitive labs can also be used to generate estimates of those patterns. During both pilot testing and final deployment, large deviations from the expected patterns would then be flagged for investigation.

After the completion of a test administration, process data can be used to understand performance in greater detail than is possible through test scores alone. Such data can, for example, illuminate group differences in scores (e.g., Bennett et al., 2021; Guo et al., 2019; Zhang, Bennett et al., 2019). Greiff et al. (2015) used process data from PISA TBA items to uncover inquiry strategy use and differential strategy use by country. Such understanding might lead to adjustments to scoring practices (when it is found that scoring privileges one type of solution), teaching (when important solution approaches are not being taught), or instruction (when it is found that a population group is not benefiting sufficiently from existing teaching practices).

At the individual level, process data could be used for formative feedback to the teacher or test taker. Those data might be used to identify test takers who are likely to be guessing (Lee & Jia, 2014), as well as those who appear to hold particular misconceptions or procedural weaknesses. In a more open-ended problem space, the level of depth at which an individual is engaging might also be described. Such evidence might enable identification of not only those who need additional guidance, but also those who could benefit from further challenge.

Analysis and Scoring of Response Process Data

The simplest process data consist of response times augmenting the raw item response. Thus, for every item, the response, its score, and the total time taken are available for analysis. Models of test speededness (van der Linden, 2017) that account for student changes in strategy due to time constraints and models that predict guessing and cheating behaviors have been applied to such data (Guo et al., 2016; van der Linden & Lewis, 2015). The remainder of this section will deal with more complex-response interactions, leading to more data available for each item response.

The rich and varied information contained within complex-response process data presents a significant challenge for analysis. The log file contains an abundance of low-level events that do not readily translate into relevant inferences. Consequently, the statistical methodology used to make those inferences needs to be well understood and appropriately validated. Here, we discuss a few of the more common approaches, along with recommendations for analysis and modeling of process data from complex tasks.

Because of the scale and complexity of the data, psychometric methods traditionally used for the analysis of item responses and test scores are frequently inappropriate. Process data are irregular in that the recorded events not only vary in number and meaning across students but also are context dependent. Furthermore, the parameter space of models used in the analysis of such data is large and assumptions of conditional statistical independence are clearly violated. For these reasons, new methods from machine learning or computational statistics may be better suited to these data. Rather than relying on models that predefine the relationships among variables, in machine learning these relationships are derived from the data. This derivation, or “learning,” requires a large amount of data; depending on the complexity of the algorithm and the number of parameters, data requirements range from thousands to tens of thousands of records. These methods are extremely useful when there is a large quantity of factors (variables or parameters) or when the relationships between the factors are complex and ill-defined. In either case, predefining a full model is impractical.

Machine learning methods are either *supervised*, which means that a labeled data set is used to train the models, or *unsupervised*, in which patterns are identified within the data without prior labeling. Labeling presents a particular difficulty for educational applications. While unambiguous classification may be more typical of some traditional machine learning domains, such as computer vision (either the picture contains a cat or it does not), “ground truth” is less common in educational and psychological

assessment. Human raters are frequently used to label data, but allowances should be made for their known fallibility (e.g., see Ho & Kane, 2013, with respect to the rating of teaching processes). Performance replays can be constructed from the process data to aid in human labeling (R. Baker & de Carvalho, 2008), which may improve consistency. Unsupervised learning is used primarily in exploratory analyses as a method for identifying common patterns within a large data set, frequently involving clustering or dimensionality reduction. For assessment purposes, such analyses can be particularly helpful in discovering response processes that are different from the expected approaches.

Both types of machine learning expect input data in which each record is represented as a feature vector—that is, a list of variables having numeric values (see also Shermis et al., this volume). The mapping of raw data to feature vectors, known as feature engineering, is a critical step because anything not encoded into the feature vector will not be usable for classification or clustering. For example, the content of text documents is frequently modeled as a “bag of words” in which the feature vector is simply the count of each dictionary word used in the document. This representation does not include word order, encoding “house boat” identically to “boat house.” The representation vector could be expanded by adding bigrams (two-word sequences) or part-of-speech tags to enable such distinctions. For analyzing response process data, features can include the count of specific actions taken, the mean time between actions, or the most frequent action in a given time slice. Feature detectors can be crafted that identify significant patterns of action within the raw data, which can then be added to the feature vector. Once the feature vectors are constructed, machine learning uses a variety of statistical techniques to classify or cluster records.

For assessment, classification can be used to evaluate performance, with categories like “high,” “moderate,” and “low.” Classification can also be used to identify records that are likely the result of guessing or cheating behaviors. For generating formative information from a summative assessment, classification can be used to identify strategies or misconceptions. Machine learning methods applied to the classification of process data include support vector machines, *K*-means clustering, logistic regression, classification and regression trees, and deep neural networks (Baradwaj & Pal, 2011; Rivas et al., 2019).

Common to all machine learning is that few assumptions are imposed about the relationship between the features in the data and the final classifications. The statistical methods iterate to optimize a loss function (e.g., classification match with prior labels, or data fit), but may add and delete factors and relationships between factors or combine the results from multiple models in a weighted fashion. This methodology makes final classifications hard to defend because the logic behind the classification is not transparent.

An alternative, or complementary, approach is theory-driven modeling. In these methods, the relationships between the data and the inference are defined in advance. The models may contain parameters that will be tuned given the data, but the meaning

of the classifications or inferences are clear *ab initio*. Bayesian networks are an example of modeling complex data with a theory-driven approach. Relationships between the features (observable variables) and the targets of inference (unobservable variables, or latent nodes) are defined by the network structure, while the exact probabilistic relationships between nodes are learned from the data.

More commonly in educational assessment, a combination of the above two approaches is used. Data mining can identify clusters of common action patterns. Content experts then examine these patterns and infer, for example, the strategy that the test takers employed. Once a set of interesting strategies is identified, models can be built to classify records by strategy or score the performance based on the action patterns observed.

One of the more active areas of process data research has been in writing, particularly for essays composed as part of standardized assessment. Using a combination of theory-driven and bottom-up approaches, this research has found meaningful relationships between essay scores and such basic features as the types of pauses that characterize composition and the length of writing bursts (Almond et al., 2012; Guo et al., 2018; Zhang & Deane, 2015). Theoretically predictable differences in feature patterns among writing task types have also been detected (Deane et al., 2018). Studies have used unsupervised data reduction methods like exploratory factor analysis to select and aggregate low-level log file features into scales (Deane, 2014; Zhang & Deane, 2015) and profiles (Bennett et al., 2022). Meaningful differences among writing proficiency levels and among gender, socioeconomic status, and racial/ethnic groups have been discovered using such scales (Bennett et al., 2020, 2021; Guo et al., 2019; Zhang, Bennett, et al., 2019).

For an extended interactive performance, a different approach from modeling a set of extracted features is to model the behaviors of the individuals within the context of the problem that they are solving. Decision models calculate the probability of a person making a choice in a particular situation, given the person's goals and beliefs (C. L. Baker et al., 2011). One can think of this approach as if we were programming an autonomous agent to perform the task. Given a goal, the agent will need to select actions, monitor the results of those actions, and select next steps until the goal is met or it gives up. Partially observable Markov decision processes are one example that can be applied to such assessment performances (Bellman, 1957; Howard, 1960). This model calculates the probability of a person taking a given action in a particular state of the problem as a soft-max¹ over the expected total rewards for taking that action. Goals are encoded into the reward structure, which quantifies both the rewards for reaching different problem states (e.g., a solution to the problem) and the costs of taking specific actions. Beliefs are encoded into the model's transition functions and state space as subjective understanding of both the probabilistic effects of taking particular actions and what is possible. Inferences about the test taker's abilities, goals, and beliefs can be made by fitting the model to the response data produced by the test taker (LaMar, 2018).

Automatically Scoring Complex Constructed-Response Tasks

The preceding section centered on evaluating response processes, a method that could be applied to a wide range of test item types and used for a variety of purposes that do not necessarily factor into the scores reported on consequential tests. In the current section, we focus on scoring *per se* and on a select class of assessment tasks. In particular, the section addresses automatically scoring a variety of complex constructed-response tasks requiring judgment of product features, process features, or both. We describe the tasks to which such artificial intelligence (AI) approaches have been commonly applied, give a high-level description of how scoring works, and suggest the types of evidence that ought to be considered for validation (see also Shermis et al., this volume).

Bennett and Zhang (2016, p. 142) offered the following definition for automated scoring: “the machine grading of constructed responses that are generally not amenable to exact-matching approaches because the specific form(s) and/or content of the correct answer(s) are not known in advance.” As they noted, that definition is quite broad, encompassing grading approaches that differ considerably as a function of the constructed-response task being posed and the character of the answers expected from a given population of test takers.

As of this writing, automated scoring is used operationally by many testing programs, including for postsecondary admissions (GRE General Test Analytical Writing Assessment, TOEFL iBT, Pearson Test of English), occupational and professional licensure (USMLE), and school accountability (selected Smarter Balanced states). The primary motivations are to reduce the cost associated with human scoring and increase the speed of reporting.

The types of tasks to which automated scoring has been applied operationally include essay writing, speaking, architectural design, patient management, accounting, mathematical problem-solving, and relatively short text responses associated with reading a passage or justifying a mathematical problem solution (Williamson et al., 2012). One important dimension along which such tasks may vary is in being static versus dynamic. For instance, in the GRE General Test Analytical Writing Assessment, the product or outcome—that is, the submitted essay response—is the only aspect graded. In contrast, the USMLE includes a section containing 13 computer-based case simulations (USMLE, 2018). As mentioned, each simulation presents a patient management problem that changes as the test taker interacts with it (e.g., the test taker’s decision to run a diagnostic test produces a result that must be considered and acted on). Consequently, USMLE automated grading must account for the *process* used to manage the patient, as well as such outcomes as the final diagnosis and prescribed treatment.²

Irrespective of the task and AI approach, automated scoring generally includes three conceptually separable parts: feature extraction, feature evaluation, and evidence accumulation (Drasgow et al., 2006). In *feature extraction*, the scorable components of the response are computed (e.g., parsing and tagging words for essay scoring; identifying what actions were—and were not—taken in managing a patient). *Feature evaluation* entails judging the extracted components (e.g., the agreement of subjects and verbs;

the appropriateness and order of actions in patient management). Finally, *evidence accumulation* involves aggregating the feature evaluations to produce one or more scores. For essay evaluation, judgments about subject–verb agreement would be combined with those related to other aspects of essay quality (e.g., organization, development, content); for patient management, evaluations of the appropriateness and order of actions would be combined with those pertaining to the suitability of the diagnosis and treatment.

Within a task type, automated scoring systems can be differentiated along multiple dimensions. One important dimension is whether scores are created to predict human scores empirically or, alternatively, are generated from theoretical propositions or rules. Most approaches to automated essay scoring take the first path. That is, they employ feature weights empirically derived to predict human scores, for example, by linearly regressing those scores on the extracted features (e.g., Burstein et al., 2013). This focus on predicting human scores has been driven by a long tradition of human essay rating, whereby human scores have come to be accepted by many users as a “gold standard” (Powers et al., 2015). An alternative approach used in some scoring systems is to employ propositions or rules based on some theoretical decomposition of what constitutes a quality performance in that task domain. The decomposition is based on expert judgment and may involve having one committee develop the rules and another committee verify the rules and the scores those rules produce, without ever optimizing the automated algorithm to predict human scores per se. That approach was essentially followed in the automated scoring of architectural design problems (Bejar, 1991; Braun et al., 2006) and in medical patient management (Clauser et al., 2016). In such approaches, humans are used as experts in defining scoring features and aggregation rules, as well as for quality control when the scores are produced.

A second dimension along which scoring approaches for a given task type may differ is specific to those approaches that seek to maximize agreement with some criterion, like human scores. The dimension is the extent to which transparent versus black-box machine learning methods are utilized. For example, in some approaches, computable features are developed from a construct theory or from an existing scoring rubric. Once computed, those features are combined by weighting them empirically to produce a score (or other judgment such as a diagnostic categorization). Divulging the computable features and their relative weights permits users to make an evaluation of the extent to which those features cover the construct theory or rubric, aspects in which the features may fall short, and how the weighting comports with the intended construct. In contrast, other approaches to automated scoring extract large numbers of computable features without any prior construct or rubric mapping and algorithmically use whatever features best predict the chosen criterion (usually human scores).³ Recent uses of large language models (LLMs) for essay scoring would appear to work in this way, though there may be ways to reduce the black-box problem via combination with more interpretable measures (e.g., Mizumoto & Eguchi, 2023). In general, however,

such approaches will not be transparent to users and may not even be scrutable to the system's developers.

The most common approach to evaluating the quality of automated scoring has been to compare the generated scores to those produced by operational human raters under the supposition that machine–human agreement equal to or greater than human–human agreement constitutes validation. As multiple commentators have pointed out, machine–human agreement is best viewed as *one* piece of evidence in the validity argument for automated scores or, said another way, validation should be broadly based (Attali, 2007; Bennett & Zhang, 2016; ETS, 2021; Williamson et al., 2012).

A considerably more comprehensive validation conceptualization has been offered by Bennett and Zhang (2016), which considers multiple sources of evidence. Their conceptualization began with the validity argument for *human* scores. It is necessary to evaluate that argument if human scores are to be used as evidence for validating the automated scores (Bejar, 2012). With respect to human scores, they asked:

- Do the test-taker response processes align with the construct definition? For example, if the computer interface is an unfamiliar one, test-taker time and cognitive resources may be split between figuring out the interface and completing the task, thereby contaminating human scores with irrelevant variance.
- Does the human scoring rubric fully capture the construct definition? If the rubric unintentionally drives human raters toward a subset of that definition, then that subset will dominate scores.
- Are operational human raters using construct-relevant scoring processes? If raters are using shortcuts (e.g., avoiding the extremes of the score scale, using correlates such as response length), then their scores will be a less meaningful evaluation criterion.
- Do raters agree reasonably highly with one another? If not, their justification as a validation criterion will be undermined.
- Do raters treat unusual responses in appropriate ways (e.g., responses that may not fit the existing rubric but clearly indicate a high level of competency; ones that attempt to game their way into a higher score)?
- Do human ratings of one task predict ratings on other tasks from the same universe reasonably well? If not, the argument for an underlying construct will be called into question.
- Do the ratings relate in theoretically predictable ways to other measures of the same construct and to measures of different constructs?
- Do the above results hold to reasonably similar degrees across important population groups? If not, those differences may reflect unfairness in human scoring.

With respect to the validity argument for automated scores, Bennett and Zhang (2016) asked:

- Was the model trained and calibrated on an appropriate sample of artifacts from the target population? If not, it may encounter responses that it cannot properly evaluate.
- Are the model's features related to one another empirically in theoretically meaningful ways and do the features and their weighting fully capture the rubric and construct definition? This question centers on whether the resulting scores are a faithful measure of the intended construct rather than some subset of it or simply a measure of one of that construct's correlates.
- Do the automated scores agree with human ratings (in the best case, with the mean rating taken across multiple experts grading under ideal conditions who agree highly among themselves)? Machine agreement with a consensus among experts makes for a more reliable and arguably more valid criterion than agreement with a single rating generated under more rushed, operational conditions.
- How effectively does the automated scoring handle unusual responses?
- How well do the automated scores predict performance on other tasks from the universe?
- Do the scores relate to external criteria in the expected ways?
- Are the functional characteristics described above invariant across population groups?⁴
- What are the likely intended and unintended impacts of using automated scoring on the behavior of test takers and those who educate them? For example, do students indiscriminately use more low-frequency words and complex sentences (when simpler vocabulary and constructions might better serve given writing purposes) because they believe such use will increase their automated scores?
- How does automated scoring compare on each of these dimensions to human scoring? Notable differences in functioning between the methods should be investigated and explained because they are likely to point to sources of inaccuracy, irrelevant variance, or unfairness in one or the other method.

In closing this section, we offer several important points. The first point is that it is generally not the automated scoring engine being evaluated, but the scores it produces. Those scores depend on the nature of the examination questions posed and the population assessed. The validity of scores may vary to the extent that either questions or population characteristics change.

A second point is that, as noted, agreement with human ratings is questionable as the sole criterion for automated score validation. This statement especially holds if the validity argument for the human ratings themselves has not been firmly established. Rather, multiple sources of evidence should be sought to permit a more rigorous and complete evaluation of the validity argument for automated scores. To the extent that those evidentiary sources suggest the same positive conclusion, the argument will be strengthened.

Third, one of those sources of evidence should be an analysis of the features used in scoring, their weighting, and the alignment of these features and weights to the construct definition. In other words, it is difficult, if not impossible, to rigorously evaluate the validity of automated scores without a clear understanding of how those scores are generated and how that generation method comports with the intended construct. Such an understanding presumes transparency of the automated scoring system. The need for such transparency is being increasingly recognized in legal frameworks governing AI systems used more generally to make decisions that have significant impact on people's lives (G20, 2019; Madiega, 2019; OECD, 2019). That recognition is, in turn, fueling efforts to create explainable AI (Dwivedi et al., 2023; Futia & Vetro, 2019; Kuang, 2017; Maglieri & Comande, 2017; Turek, n. d.).

A final point is that our efforts to build a strong validity argument for constructed-response scoring, whether automated or human, should ramp up as the consequences associated with test results increase. This stipulation holds even when those constructed-response scores are not the major portion of the test. Such is the case because, when the decisions emanating from test results have significant and hard-to-reverse impact, it is the testing program's responsibility to ensure that, to the maximum extent practicable, all test components are meaningful and fair indicators of proficiency.

ASSESSMENTS USED TO SUPPORT INSTRUCTIONAL DECISION-MAKING

The section "Assessments Used to Support Consequential Purposes" dealt with consequential TBA, where decisions based on a single result may have a dramatic influence on an individual, group, or institution and may not be easily reversed. In the current section, we focus on TBA uses that generate more easily reversible decisions with less dramatic effect in any given instance. In particular, we discuss the current landscape for TBAs used to support instructional decision-making in real time at the individual student level. We will briefly trace some of the important milestones in the development of such embedded, technology-based formative assessment, noting relevant innovations in design, task type, scoring, and modeling.⁵ (See also Brookhart & DePascale, this volume, for a discussion of formative assessment.)

The use of technology-based assessment for instructional decision-making is premised on the principle that individualizing instruction, as would be done in one-to-one tutoring, is more effective than targeting instruction to the group. This principle was implemented through mechanical devices known as teaching machines, which used simple forms of assessment to direct instruction. One of the earliest such machines was created by Pressey (1926, 1927), whose initial purpose was to build a testing device for presenting and scoring responses to multiple-choice questions. Realizing its value for instruction, he incorporated such rudimentary mastery criteria as not eliminating a question until it had been answered twice correctly. Following publication of Skinner's (1958) seminal "Teaching Machines" article, the same basic ideas led to CAI, which

afforded more complex branching possibilities based on the student's prior response(s) (Suppes, 1972; Van Meer, 2003).

The 1980s saw the advent of intelligent tutoring systems (Sleeman & Brown, 1982), today more commonly called adaptive learning or personalized learning systems. Whereas teaching machines, and the CAI implementations that succeeded them, were built on behaviorist conceptions of learning, intelligent tutors brought cognitive learning models to bear and combined them with AI approaches (Anderson et al., 1985). Those domain-specific cognitive models and AI methods were used to track a student's knowledge state dynamically as it evolved during an instructional sequence and progressively adjust instruction as a function of that changing knowledge state. This real-time assessment was initially deterministic, meaning that uncertainty was not factored into the evaluation of the student's state. These deterministic systems were followed by ones that incorporated probabilistic modeling, typically in the form of Bayesian networks, into their real-time assessment of student knowledge state (Corbett & Anderson, 1995; Mislevy & Gitomer, 1995; VanLehn & Martin, 1998).

Many of the intelligent tutors created over the past several decades did not go beyond the prototype development stage and, consequently, were not widely used in classroom settings (Shute & Zapata-Rivera, 2010). One tutor that has been widely used is Carnegie Learning's MATHia, a direct descendant of Anderson and colleagues' extensive, long-term research program at Carnegie Mellon University (Anderson et al., 1985, 1995; Pane et al., 2013; Ritter et al., 2007). A second widely used tutor is ALEKS (Assessment and Learning in Knowledge Spaces), based on the research of Falmagne and associates at the University of California, Irvine (Doignon & Falmagne, 1999; Falmagne et al., 1990). Both ALEKS, now a product of McGraw-Hill Education, and MATHia are used at the school level as well as in higher education. A third, more recently developed example is Woot Math, which helps students in Grades 3–8 learn core math concepts, starting with rational numbers (Milne et al., n.d.). Rather than the more elaborate cognitive-domain modeling approach taken in MATHia, Woot Math concentrates selectively on a small number of ideas and misconceptions identified as key by expert teachers and researchers. Of note is that, in addition to selecting tasks based in part on Rasch models, the system adaptively determines how fast each student should move along the instructional sequence, offers help based on Bayesian models of each student's understanding, presents additional instructional modules, and inserts new levels with review tasks.

A final example can be drawn from the class of tutors known as personalized learning apps, which can be accessed on mobile phones, tablets, or conventional computers. As an instance from this class, the Duolingo language learning app (<https://www.duolingo.com/>) offered as of this writing instruction in three dozen or so languages, with the more common language courses having on the order of 150–200 brief lessons. Lessons are organized around topics common to language learning (e.g., shopping, food, entertainment), into which new vocabulary and linguistic structures are progressively integrated. Each lesson is composed of an optional synopsis of the content to

be covered and exercises taking various forms (e.g., matching, selecting and arranging words to form a sentence, speaking a presented sentence). Adaptivity and such learning sciences' principles as spaced repetition are employed in exercise selection via probabilistic models (Protalinski, 2020; Settles & Meeder, 2016), although the particular models used operationally have not been identified.

Intelligent tutors, and their CAI and teaching machine predecessors, had in common that instruction was generally built around the presentation and solution of a series of discrete problems designed to evaluate student standing. In contrast, educational games and simulations often present a more thematically integrated problem-solving experience as a means of facilitating and monitoring learning. That experience may require more problem-solving steps, and the state of the game or simulation may change in response to student actions. Thus, educational software based on games, as well as on simulations, is often more performance oriented.

Like intelligent tutors, games and instructional simulations can, in principle, be built on cognitive-domain models, use AI, and incorporate probabilistic methods to estimate skill level. Estimation can be dynamic, allowing real-time modulation of the state of the problem situation in play or of the difficulty level of the problem situation to be presented next. As of this writing, we could locate no commercially available, widely used, and well-documented examples that possessed all these features. Several commercially available games and simulations, however, do possess one or more of these attributes.

As an example, Math Garden allows students to practice basic addition, subtraction, multiplication, and division skills (Klinkenberg et al., 2011). The game is in many ways quite traditional, presenting a series of drill-type math problems for which correct solutions add garden flowers and garner coins for buying virtual prizes. In addition, no cognitive domain model or AI is employed. Presentation is adapted at the item level using estimates of student competency that incorporate both speed and accuracy. Estimates are dynamically generated via the Elo (1978) rating system (originally developed for chess competitions).

An example from science is Inq-ITS (Gobert et al., 2018; <https://www.inqits.com/>). This system employs a collection of laboratory simulations that allows students to design and conduct experiments. Inq-ITS uses the Next Generation Science Standards as its cognitive-domain framework, coaching students through the NGSS practices of designing investigations, using evidence to make claims, and backing up those claims with evidence and reasoning. The system employs machine learning techniques to evaluate student responses, generate real-time alerts for teachers, and give students feedback on specific aspects of their inquiry practice. Bayesian knowledge tracing, described in the next section, is employed to estimate student skill levels (LaMar et al., 2017, pp. 143–145).

The formative assessment designs, tasks, scoring, and modeling methods used in intelligent tutors, games, and simulations can run from simple to complex. Common to most systems of interest are presenting a sequence of problems created to facilitate learning some set of domain competencies, scoring in real time, probabilistically

estimating knowledge state (or skill level) from that scored performance, and using that estimate to in some way modulate the current task or choice of the subsequent one. In the following paragraphs, we briefly discuss similarities and differences among these systems in the problems employed, presentation design, response requirements and evaluation, and feedback.

As noted, the problems that these systems present may be discrete or part of a thematically related sequence. Thematically related sequences are often introduced by a scenario intended to give a real-world context, motivate the student, and indicate a goal to be achieved. One or more problems typically follow. In an intelligent tutor, a problem statement, related stimuli, and any associated tools are available to the student. A game would add mechanics (i.e., rules and procedures for play) and such motivating elements as points for correct responding, graphics, and audio. In educational simulations, the student interacts with an invented milieu that is intended to represent the key features of some real-world environment, such that the invented environment mimics the real one. The simulated environment may include runnable models that process inputs and produce outputs like their real-world counterparts (e.g., behaving like a patient with a specific medical condition). Problems may also be presented outside the simulated environment, in which case the student responds to those problems based on his or her interactions with the simulation. Alternatively, or in addition, the students' interactions with the simulation may constitute evidence for evaluating knowledge state. Regardless, the simulated environment itself becomes part of the problem.

With respect to presentation, in intelligent tutors, presentation is typically adaptive. Games and simulations, however, may have linear, item-level adaptive, or multistage adaptive designs. In linear designs, all students encounter the same problems in the same order. Many educational games take this approach, allowing a student to progress to a higher level only when some criterion performance has been reached (e.g., a fixed number of problems answered correctly). In item-level adaptive designs, the difficulty of the next problem is determined at least in part by performance on earlier problems. Multistage designs differ from item-level adaptive ones primarily in the frequency of adaptation, adjusting the difficulty of tasks at the stage level, rather than at the item level.

As to response requirements and evaluation, in intelligent tutors, games, and simulations, such requirements may involve entering a number, clicking on a hot spot, moving objects on screen, writing text, or manipulating sliders, dials, or other components. Responses in all three types of systems can be evaluated in terms of the resulting product, the process used to generate that product, or both, depending on the underlying domain model. However, even when the domain model specifies that evidence of proficiency lies in the correctness of an outcome (e.g., a mathematical result), examining the process used to generate that outcome can be the basis for action. That action could take the form of feedback (e.g., providing hints, pointing out errors in process, showing a worked example), choosing new problems, or making other adjustments to instruction.

Feedback itself may be provided before or after responding. A response that is not given quickly enough may trigger some event to occur. That event could be visible, as in the provision of a hint, or invisible, as in updating the estimation of the student's knowledge state. Feedback may also immediately and visibly follow an action—for example, an English-language-learning game character might offer a puzzled look if a grammatically incorrect sentence is entered.

Probabilistic Models for Assessment Embedded in Instructional and Learning Systems

As indicated in the section immediately above, to provide feedback, as well as to adjust instruction, the formative assessment embedded in the learning and instructional systems of interest uses probabilistic models. In this section, we explore some of these models, their evaluation, and their validation.

Within tutoring systems, “student models” are used to track relevant parameters for various characteristics of interest including proficiency, engagement, and affective state (Johns & Woolf, 2006). These models are essential for enabling personalization of the tutoring experience, with the most basic model indexing where the student might be in relation to mastering the content being taught. In such models, proficiency is estimated dynamically from the student's performance to enable real-time adjustment to instruction. Estimation is done through some form of psychometric (or other probabilistic) model.

Although adaptive tests also dynamically estimate proficiency, they presume no change in proficiency over the course of the assessment. In contrast, proficiency should be expected to change because of interacting with a tutoring system. In addition, the system must be able to generate frequent formative feedback. Effective feedback requires specific, accurate, actionable observations about student performance (Hattie & Timperly, 2007; Shute, 2008). Thus, models need to be able to estimate specific strengths, weaknesses, and, in some domains, misconceptions.

For adaptive learning systems like intelligent tutors, the purpose of the student model extends beyond direct feedback to enabling the selection of appropriate next-instructional steps. These interventions might be in the form of hints, encouragement, explanatory text or video, a worked example, or selection of the next problem.

At a high level, the student model can be decomposed into the representation of the student's state and the method used for updating that representation. Approaches to representing student state vary considerably, but some common ones include overlay models, which characterize the student's knowledge as a subset of an expert's knowledge; perturbation (or buggy) models, which catalog both the student's correct and incorrect ideas or misconceptions; and stereotype models, which represent the student through membership in predefined classes (Chrysafiadi & Virvou, 2013; Desmarais & Baker, 2012). The methods for estimating and updating these models also vary widely, but frequently they involve a combination of a hidden Markov model (HMM) and a more traditional psychometric model.

The dynamic nature of updating estimates of student state makes the HMM a particularly suitable statistical tool. The HMM (see Figure 9.5) specifies how an unobserved latent variable, for example, mastery of skill K , might change over time based on observed variables, x_i , for example, answers given to problems within the learning system. The statistical relationship between the unobserved variable and the observed one is known as the emission probability, $p(x_i = 1 | K_i)$. For many student models, this emission probability takes the form of a psychometric model (e.g., IRT, cognitive diagnostic model, CDM; Rupp et al., 2010). The probability of the latent variable changing between student actions is known as the transition probability, $p(K_i = 1 | K_{i-1})$. Within adaptive learning systems, this probability is related to the learning rate. Additional parameters, frequently represented as ϵ , can include amount of time between occurrences and a quantification of the intervening instruction.

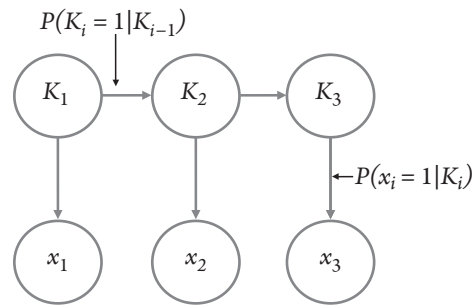


FIGURE 9.5
A Hidden Markov Model for Estimating Student State Over Time

Because the form and content of adaptive learning systems vary widely, the modeling approaches also vary. For learning systems that primarily present a series of problems for the student to work through (e.g., ASSISTments, Heffernan & Heffernan, 2014; Andes, vanLehn et al., 2005), the key inference at any given moment is whether the student has mastered a particular, narrowly defined skill. This inference allows the tutor to introduce a new skill once the current one has been mastered. Thus, an overlay model is frequently employed to represent the student, with a discrete set of Boolean variables tracking mastery of the skills or knowledge components of interest. The most common updating method for this purpose is Bayesian knowledge tracing (BKT; Corbett & Anderson, 1995), which takes the form of an HMM. In BKT, the probability of a student having mastered a skill L on their n th attempt is modeled as the sum of the probability they had previously mastered it at the $(n - 1)$ th attempt and the probability that they had not previously mastered it but now have mastered it (transitioned) on this attempt (T_n):

$$p(L_n) = p(L_{n-1} | X_{n-1}) + (1 - p(L_{n-1} | X_{n-1})) * p(T_n) \quad (1)$$

The emission probabilities in BKT usually take the form of a CDM (Rupp et al., 2010), which models correctness of response based on a set of skills represented as Booleans, either mastered or nonmastered. CDM models include slipping and guessing parameters to account for an incorrect response when the requisite skills are present (slipping) and a correct response when the requisite skills have not been mastered (guessing). Combinations of skills can be modeled either as conjunctive or as compensatory, allowing the flexibility to apply appropriate cognitive theory for the domain.

Another common approach to student modeling within adaptive learning systems is performance factor analysis (Pavlik et al., 2009). Such models characterize student knowledge in terms of continuous variables and make use of an updating mechanism that is a variation of IRT with a learning parameter (slope).

More recently, advances in recurrent neural networks have been applied to the problem of estimating and updating student proficiency variables in what is known as deep knowledge tracing (DKT; Piech et al., 2015). DKT uses the same inputs and outputs as BKT, but rather than computing probabilities of skill mastery and using them to estimate subsequent performance, the probabilities of success on future items are predicted directly with a recurrent neural network, usually using a long short-term memory layer to ensure that past performance has an extended impact on future predictions. At the time of this writing, several extensions to DKT have been developed that augment the input vectors with both item information (including difficulty and skill tapped) and additional performance information, such as the amount of time taken to solve the item (Ai et al., 2019). Structural improvements to the predictive models have also been proposed, including applying regularization to prevent large variance in the predictive output (Yeung & Yeung, 2018). While the recurrent neural networks approach requires a large training data set, the flexibility of these models to utilize complex inputs, incorporate domain-specific data transformations, and interface with other models makes it likely that DKT and other uses of deep neural networks will have a major impact on adaptive learning technology in the near future.

Models that are more deeply grounded in learning theory have also proven useful for interpreting behavior in solving complex tasks. In particular, the ACT-R cognitive architecture (Anderson et al., 1997) has been used to predict student actions in multistep problem-solving using a technique known as *model tracing* (Corbett et al., 1995). ACT-R models cognition at a fairly low level, including declarative memory, working memory, and procedural knowledge in the form of if-then production rules. These models predict not only keystroke-level student actions, but also the time between actions, providing additional predictions that can be used for model validation. Because there are so many parameters involved in an ACT-R model, individual parameter estimation is not attempted. Instead, student actions are compared to an ideal student, or expert, model. In some adaptive learning systems, this comparison is sufficient because any off-path action triggers tutoring until the student returns to the ideal path.

Bayesian models and estimation have been popular in adaptive learning systems, including the intelligent tutors built on ACT-R (Koedinger & Corbett, 2006). One

attraction is that Bayesian estimation can occur iteratively as each new data point updates the models' probability distributions. This real-time updating allows the system to act on an estimate as soon as it passes a set certainty threshold. In ACT-R tutors, this "knowledge-tracing" technique allows the system to select tasks that exercise knowledge components the student is not likely to have already mastered (Alevan et al., 2017).

One particular type of Bayesian model, the Bayesian network (Bayes nets), is frequently used to represent complex relationships between the evidence emerging from student performance and the latent variables that make up the student model (Chrysafiadi & Virvou, 2013). A Bayes net is a directed acyclic graph in which nodes represent variables and relationships between nodes are described with conditional probability distributions. In adaptive learning systems, the variables represented are almost always categorical, making these relationships conditional probability tables (CPT).

Figure 9.6 shows a simple Bayes net that might be part of a psychometric model (note that there are too few observable variables for this model to be identifiable given the number of latent variables). The observable variables (x_1, x_2, x_3) are indicators (behaviors) from a student performance, while K_1 and K_2 are latent variables that explain differences in performance. In this example, we include an intermediary latent node S_1 that could indicate a particular strategy students might implement. The observable behaviors x_1 and x_2 are indicators of use of this strategy and, as such, are not conditionally independent given K_1 , but are conditionally independent given S_1 . Observable x_3 has two parents (K_1, K_2) making this variable a within-observable multidimensional measurement model. The way in which K_1 and K_2 interact to produce the observable x_3 can be flexibly defined by the CPT of x_3 , allowing for compensatory, noncompensatory, and more complex relationships to be modeled.

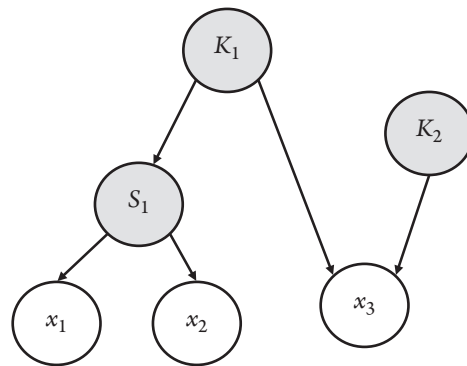


FIGURE 9.6
A Simple Bayesian Network

Note. The gray nodes represent latent variables and the white nodes are observed variables.

A Bayes net can be updated when new observations are made, giving a running dynamic probability distribution for the latent variables. One of the most attractive properties of using Bayes nets for student models is that complex relationships between cognitive and affective latent variables can be modeled to explain specific behaviors. Thus, for learning systems that employ dynamic estimation from student interactions in a simulation, game, or other complex performance, the Bayes net is a natural choice. These models can also be found in formative assessments that include both complex evidence and student models (Almond et al., 2007, 2015).

Evaluation of Student Models in Instructional and Learning Systems

The overall functioning of adaptive learning systems has generally been evaluated based on the achievement gains produced as compared to a traditional curriculum (e.g., Pane et al., 2013). While such gains may suggest that the system is working properly, they do not necessarily confirm the validity of the student models. Similarly, the absence of learning gains does not necessarily imply a failure of the models. Because the function of the student model is to enable feedback or adaptation of tutoring, the model's value might best be judged with respect to how effectively it achieves these goals during the learning session. Experimentally contrasting a linear presentation of problems with one dictated by knowledge tracing, for example, can identify the student model's impact on achievement (Corbett et al., 2000, as cited in Aleven et al., 2017, p. 528).

Other evidence as to the quality of model functioning can be provided by how well the models predict the student's next action or the success of their next solution attempt. Because this prediction is seen as a classification task, researchers in the learning system and data-mining communities have frequently used a classifier accuracy metric known as area under the curve. Area under the curve is calculated by plotting the probability of a true positive by the probability of a false positive over the range of possible threshold values for the classifier. The area under the curve would thus be 1.0 for a perfect classifier and 0.5 for random selection.

COMBINING CONSEQUENTIAL DECISION-MAKING AND INSTRUCTIONAL SUPPORT: CAN TECHNOLOGY-BASED ASSESSMENT DO BOTH?

Earlier sections addressed assessments used to support consequential purposes and assessments used to support instructional decision-making. The affordances of technology make it easier to imagine the possibility of generating information from the same assessment that might simultaneously serve consequential decision-making and instructional-support purposes. Several approaches might be taken to satisfy this dual end, with the approaches differing in fundamental ways, especially in foregrounding one or the other purpose.

The least radical approach would entail generating formative information from consequential TBA. Here, the consequential purpose would be central and the formative purpose supplementary. As an example, a state English language arts accountability assessment will typically produce total scores, proficiency levels, and other scores. One or more of the other scores may come from a response to an essay task, including scores summarizing overall writing quality as well as ones characterizing such writing traits as content and conventions. To be sure, such trait information provides only a very minimal level of formative information. More valuable might be a description of *how* the student wrote the essay, generated by capturing and analyzing the keystrokes used to create it (Zhang, Bennett, et al., 2019). For instance, was there evidence of planning (e.g., the rudiments of an outline early in the writing session), verbal fluency (e.g., long bursts of text), word-level monitoring (e.g., correction of typos), or global monitoring (jumps from one cursor position to another followed by insertions and deletions)? Were there indications of difficulty in typing (signaled by consistently lengthy intervals between key presses) or of problems in word finding (e.g., long pauses between words)? Possibly more valuable for the teacher and student would be the ability to replay, in speeded-up time, the essay as it was written, thereby giving an opportunity for discussion and reflection on the writing process (Vandermeulen et al., 2020). A process description and replay could lead to the realization that typing practice was needed or that instruction in planning or in how to edit should be provided.

The above approach clearly foregrounds consequential assessment by building into it secondary mechanisms for more effectively supporting instructional decisions. Those mechanisms will be far from ideal. Such feedback will occur late in the school year, be limited to that which can be either wrung from the existing assessment or designed into it without compromising its primary purpose or practicality, and be divorced from the context of classroom instruction. Along with other reasons, these limitations have motivated third-generation proposals that attempt to better balance the two assessment purposes. That is, these proposals move toward foregrounding instructional support decisions, with information for consequential decisions coming as a by-product. Periodic, or through-year, assessment offers one version of this idea (Bennett, 2010b; Bennett & Gitomer, 2009; Northwest Evaluation Association, 2019).

In through-year assessment, measures are repeated at various time points. A key benefit of this model is that consequential decisions (e.g., school accountability) would no longer be based on a single measure taken at one time point, but rather on some aggregation of the through-year observations. Also, more frequent feedback would be provided from each of the through-year events.

Many features of this idea could in principle be varied, including the number of TBAs administered, what content they cover, when they are given, the extent to which districts and schools can choose the ordering, and how results are accumulated into a performance index. The greater the variation in these features across schools or districts, the lower will be the comparability of the results (Bennett, 2020). The highest level of comparability would generally result from administering a fixed number of assessments

created from the same content specifications at selected time points to each student under the same conditions and scored with the same aggregation rules. Administering different forms of a comprehensive interim assessment during 1-week windows in fall, winter, and spring would be a simple example, though the extent to which it would foreground instructional support decisions might be arguable (Shepard et al., 2011). More attuned to instructional support decisions would be a regimen in which each assessment focused on a different critical competency associated with grade-level standards. Such a design would either require all schools to follow the same instructional sequence or allow them to choose the order of administration, complicating comparability and test security. (See Georgia Department of Education, 2018, pp. 80–82, for a description of such an assessment piloted under the federal Every Student Succeeds Act Innovative Assessment Demonstration Authority.)

In through-year models, many rules for aggregating results are possible, each derived from a different conception of achievement (Wise, 2011). Weighting the results as a composite with the greatest emphasis given to the most recent assessment would reflect a view of achievement as competency accumulation. A different weighting could mimic the way grades are averaged over quizzes, midterms, and final exams, with this more even weighting suggesting a view of achievement as a collection of accomplishments. A third possibility is to aggregate results based on test information, which would emphasize precision in measuring the dimension common to the several assessments. Finally, taking some function of the difference between the first and last result conceptualizes achievement as the extent of growth. Psychometric models related to these different types of aggregation have been explored by Mislevy and Zwick (2012), Fu and colleagues (Fu & Feng, 2018; Fu et al., 2012, 2013), and Rijmen (2009).

Perhaps the most radical proposal for foregrounding instructional support decisions while generating information for consequential decisions as a by-product is to use a record of daily learning interactions, potentially removing the need for separate assessments entirely (Bennett, 1998, pp. 11–14; Gee & Shaffer, 2010; Pellegrino et al., 2001, pp. 283–287; Tucker, 2012).⁶ There are many attractions to this third-generation idea, including the sociocognitive one that the contexts for learning and assessment become identical. That is, both activities draw on the very same content, knowledge representations, and tools (Bennett, 2015), increasing the value of the derived information for adjusting instruction. An example of how this idea might work instructionally can be found among adaptive learning systems like MATHia (Ritter et al., 2007). As noted, these intelligent tutors estimate from real-time learning interactions what a student knows and can do with respect to a content domain and then base moment-to-moment adjustments on that dynamically computed estimate.

Using this type of data for consequential decision-making, however, would be very challenging (Bennett, 2015). One obvious concern is that the assessment model is essentially an extreme case of through-year assessment—that is, one without constraints. The reality is that, within a state and even within many districts, students in the same grade use *many* different electronic learning applications. These applications

vary in their content coverage, emphases, and order; knowledge representations and tools; rigor and quality of problems posed; and scoring criteria and methods, among other things. Further, some environments are likely to be employed more frequently with certain demographic groups, thereby confounding group and measure. A related, but broader concern is that the evaluation instrument is no longer independent of the school or district, leading to the perception that these entities are appraising themselves. Also possibly problematic is the continuous recording of behavior for use in consequential decision-making, which raises both privacy concerns and such educational ones as the potential for discouraging experimentation in teaching and learning. Discouraging experimentation would be unfortunate because such behavior serves a critical function in learning (Kapur, 2010).

Several possibilities exist for reconciling more effectively the goals of consequential assessment with those of assessment for instructional support. These possibilities revolve around systems of assessment—that is, TBAs designed to function synergistically in their pursuit of different goals because trying to fashion a single method for achieving competing goals leads to a suboptimal solution for each goal. One such proposal follows the competitive sports model (Bennett, 2015), a variation of through-year assessment. During instruction, students utilize whatever electronic and other learning environments their schools employ, with learning interactions recorded and used for guiding (and occasionally describing) instruction but never for consequential purposes (as in the practice periods before, and often interspersed within, a sports competition). Students and teachers are informed when an assessment for consequential purposes is to occur (i.e., the actual competition). That assessment is common in design, content specifications, administration window, and scoring across all schools and districts. For state policy makers, such a system could provide the data needed to evaluate how well individual schools were performing in educating all groups of students (from some aggregation of the common assessments' results), along with descriptive data about what students in each school were doing (from sampling the recordings of learning behavior). The latter data might allow instruction to be described at an unprecedented level of detail, greatly enhancing our understanding of learning activity, content, and rigor differences occurring across teachers, classes, schools, districts, and demographic groups.

SUMMARY AND CONCLUSION

This chapter considered three broad classes of TBA: ones used to support consequential purposes, ones used to support real-time instructional decision-making, and third-generation ones that attempt to combine both purposes.

For consequential testing programs, the rationale for moving to TBA is tripartite: align the testing medium with that of learning and of the information economy workplace, conduct assessment processes more efficiently, and measure what previously could not be measured as well or at all. In the United States, a first-generation

infrastructure exists for achieving these goals. Many state assessments, national and international group-score assessments (NAEP, PISA, Programme for the International Assessment of Adult Competencies [PIACC]), graduate and professional admissions measures (GRE General Test, TOEFL iBT), and occupational and professional assessments (USMLE, Architect Registration Examination, Praxis) are routinely delivered by computer in educational institutions or in dedicated centers. Other assessments, most prominently college entrance examinations, are beginning to be delivered in that mode as well.

As a result of the COVID-19 pandemic, some consequential programs have moved to include delivery via remote human and AI proctoring to test takers' homes and offices. Such administration had been used successfully on a small scale in a few niches for several years (e.g., competency-based education at the university level and, more recently, English language assessment via the Duolingo English Test). The COVID-19 pandemic, however, greatly accelerated testing at test-taker locations. In 2020, the GRE General Test, TOEFL iBT, Praxis, HiSet (a high school equivalency examination), the Law School Admission Test, and the International English Language Testing System added home options. Whereas there has been relatively little published research on the technical quality of assessments so delivered, we expect this evolution to continue as need drives use, with research catching up.

The successful deployment of a robust first-generation delivery infrastructure for consequential testing has offered a foundation for innovations in assessment design. Those innovations generally aim to increase construct fidelity by evoking more complex cognitive processes and allowing responses to be observed in finer detail. Four kinds of innovative assessment design were distinguished: TEIs, extended-interaction items, scenario-based tasks, and simulation-based performance tasks. Gathering validity evidence regarding the functioning of these types is critical to asserting that the evoked processes are in fact the intended ones, that the innovations do not introduce unfairness for groups or individuals, and that the Person $\times$ Task interaction associated with scenario-based and simulated-based performance tasks is appropriately accounted for.

The changes to practice that TBA entails inevitably raise challenges for score comparability. These challenges occur because many programs offer tests in both paper and technology modes, other programs wish to maintain TBA score continuity with past paper versions, and still other programs want to ensure constancy of score meaning when the digital test is offered on multiple technology platforms. Different degrees of comparability may be appropriate for different use cases and achievable through different methods. Comparability may be studied, and the data needed to make score adjustments gathered, via various designs. However, technology is changing rapidly, forcing assessment programs to change as well. That rapid evolution may make it impractical to study and adjust continuously for the effects of new assessment implementations. Thus, it may be necessary to replace an empirically focused approach with one grounded in design principles directed at maximizing validity and fairness for individuals and groups. Empirically based methods like those traditionally used for linking may still play a role

as one of the investigative methods used to derive assessment design principles or to document the degree to which device-agnostic results were obtained. However, empirical approaches cannot be relied on as the primary mechanism for ensuring and maintaining comparability. Moreover, the desire to maintain comparability over time needs to be balanced against giving testing programs license to incorporate the innovation required to remain relevant in a rapidly changing world.

One of the as-yet underutilized benefits of TBA is certainly its ability to capture evidence of response processes. Such evidence can contribute to the validity argument for an item or task, help in the identification of guessing or cheating, and indicate strategy use or the presence of misconceptions. Response process data may include actions, resulting events, and latencies. Those data may index relatively basic cognitive processes like fluency or higher level strategies, plans, and knowledge. Because of their complexity and scale, process data are often analyzed and modeled using relatively opaque, bottom-up methods like machine learning, as well as ones from computational statistics that allow for the instantiation of theoretical propositions (e.g., Bayesian networks). Combining bottom-up and theory-driven approaches holds promise in that machine learning can be employed to help build the theory to be implemented in, for example, a Bayesian network.

In contrast to response process data, the automated scoring of complex constructed responses is used operationally in many consequential testing programs. This use is driven by the need to cut costs and increase reporting speed. Such scoring usually focuses on an end product, such as an essay or an architectural design, although process data are also included in some instances (e.g., medical patient management). Multiple approaches can be used to score a given task type. Important distinctions among methods relate to whether the automated scores are created to predict human scores empirically or generated from theoretical propositions or rules. A second distinction specific to approaches that seek to maximize agreement with a criterion like human scores is the extent to which transparent versus black-box methods are employed. Whereas validation of scoring has usually focused on agreement with human scores, such agreement is best viewed as one piece of evidence in a more comprehensive validity argument. That more comprehensive argument may, among other things, need to include an evaluation of the validity of using the human scores as a criterion. Additionally, how automated scores are produced, and the alignment of that method with the intended construct, should be a consideration. Because of concerns over the use of AI in society generally, we should expect significant work on explainable, transparent scoring methods.

The chapter's second major section dealt with assessments used to support instructional decision-making. Such assessment is incorporated into intelligent tutoring (or adaptive/personalized learning) systems, educational games, and simulations. Intelligent tutors typically deliver instruction built around discrete problems, which may be similar to the TEIs and extended-interaction items found in consequential tests, whereas games and simulations often present a more thematically integrated problem-solving experience utilizing performance tasks. Tutors, games, and simulations

may employ some combination of cognitive domain model, AI, and probabilistic method to estimate competency dynamically and, thereby, modulate the current problem or choice of the subsequent one. Competency is tracked through a student model that includes a representation of the student's knowledge state (referenced to the larger cognitive domain model) and a method for updating that representation. Probabilistic methods employed in that updating have included BKT and Bayesian networks. MATHia, ALEKS, and the Duolingo language-learning apps are examples of such tutors in widespread use. Examples of games or simulations that appear to personalize learning via such methods include Math Garden and Inq-ITS.

In tutors, games, and simulations, response evaluation can focus on the product (e.g., answer to a math problem), the process used to generate it, or both, depending on the underlying cognitive domain model. However, even when the model specifies that evidence of proficiency lies in the correctness of a product, analyzing the process can suggest appropriate next steps in the form of feedback, selecting new problems, or making other instructional adjustments.

Used in some commercial instructional applications are automated scoring technologies, most commonly for writing or speaking, that share methodology with those employed for consequential assessment. Examples in composition include MyLab Writing and Criterion. These systems evolved separately from intelligent tutors. As a result, the former systems do not generally use cognitive domain models or probabilistic methods to estimate and represent student competency or to personalize instruction, instead only rating and giving feedback on each response in isolation. We should expect to see systems that use automated scoring to support instruction converge with intelligent tutors because the two approaches offer complementary capabilities.

The last section in this chapter described third-generation approaches that combine instructional support with consequential purposes. Approaches differ in the extent to which instructional support decisions or consequential decisions are foregrounded. Included were generating formative information from consequential TBA, through-year assessment, and generating consequential information from ongoing learning interactions. The last approach, while seemingly attractive, raises questions of comparability, privacy, and the potential for negative educational consequences. An approach based on the competitive sports model was proposed as a path that might offer the desired benefits while mitigating the issues raised by more radical models.

What might be productive directions for research? There are many possibilities and here we suggest but a few. One important direction for both consequential testing and instructional support purposes might be to build the foundation for operationalizing the use of process data. For example, research should be directed at identifying whether diagnostic profiles can be identified suggestive of differential instructional action. Placement in a profile could be part of the formative output from a consequential test (e.g., a writing assessment) or as part of personalized instruction. Questions concern

whether meaningful, distinct classes of behavior can be identified (e.g., dysfluent writer vs. fluent writer), whether placement of individuals in such classes is stable across tasks from the same domain, whether the addition of product information adds value, and whether instruction based on such placement results in more learning than a profile-agnostic alternative.

An associated direction concerns how best to report process data, whether in conjunction with a profile or not. If the reporting goal is to encourage reflection as to how one went about problem-solving, then performance replay should be studied. Key questions include identifying the replay formats that best facilitate reflection, learner action, and change in competency.

A third direction relates to the impact of simultaneously incorporating cognitive domain models, AI, and probabilistic methods into games and simulations. These capabilities are not yet commonly found together in commercial products. Experiments could be conducted to estimate the incremental effect of using such additions to increase personalization. A similar direction could be taken with respect to the insertion of cognitive domain models and probabilistic methods in learning systems that use automated scoring.

A fourth direction is associated with the implications of LLMs for validity, modeling, and analysis. The potential uses of LLMs are wide ranging, including item generation, constructed-response scoring, feedback, and reporting, among other possibilities (Bulut et al., 2024; Hao et al., 2024). This direction is notable because of the great interest evident in the field and the potential for improvements in efficiency and quality it seems to portend. Although this direction may pose new issues, because LLMs are a subcategory of AI, many of the challenges LLMs bring are the same ones as already noted in, for example, the section on Automatically Scoring Complex Constructed Response Tasks (e.g., bias, explainability, the need for validation criteria to be broadly based), as well as in other publications (e.g., Bejar, 2012; Bennett & Zhang, 2016).

Finally, calls to merge the purposes of consequential tests and instructional support are likely to grow as learning activity increasingly occurs online. Research should focus on studying the technical quality and instructional utility of approaches that attempt to account in principled ways for the challenges inherent in combining these divergent purposes. The competitive sports model is one example. Research should attempt to determine whether it can produce meaningful consequential results at the same time as it describes and guides instruction.

ACKNOWLEDGMENTS

We wish to acknowledge Steve Ferrara, Ryan Baker, and Denny Way for their helpful reviews on earlier drafts of the manuscript. Steve's reviews were as he was in life: constructive, gentle, modest, optimistic, and wise. This chapter is dedicated to his memory.

REFERENCES

- Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In M. Desmarais, C. F. Lynch, A. Merceron, & R. Nkambou (Eds.) *The 12th International Conference on Educational Data Mining*, (240–245). Educational Data Mining.
- Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522–560). Routledge.
- Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (ETS Research Report No. RR-12–23). ETS.
- Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, 44(4), 341–359.
- Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. Springer.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). *Intelligent tutoring systems*. *Science*, 228, 456–462.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167–207.
- Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, 12(4), 439–462.
- Association of Test Publishers. (2017). *Innovative item types: A white paper and portfolio*. Author and Institute for Credential Excellence.
- Attali, Y. (2007). *Construct validity of e-rater® in scoring TOEFL® essays* (ETS Research Report No. RR-07-21). ETS.
- Backes, B. & Cowen, J. (2018, April). *Is the pen mightier than the keyboard? The effect of online testing on student measured achievement* (Working Paper 190). American Institutes for Research.
- Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief–desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*, 33, 2469–2474.
- Baker, R., & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. *Educational Data Mining, 2008*, 2469–2474.
- Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6), 63–69. <https://arxiv.org/ftp/arxiv/papers/1201/1201.3417.pdf>
- Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology*, 76, 522–532.

- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2–9.
- Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation* (ETS. Research Memorandum No. RR-99-2). ETS.
- Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 6, 679–684.
- Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Policy Information Center, ETS. https://www.ets.org/research/policy_research_reports/pic-reinvent
- Bennett, R. E. (1999). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S. Messick (Eds.), *Assessment in higher education: Issues of access, student development, and public policy* (pp. 181–192). Lawrence Erlbaum Associates.
- Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment*, 1(1), 2–23. <http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1667>
- Bennett, R. E. (2010a). Technology for large-scale assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 8, pp. 48–55). Elsevier.
- Bennett, R. E. (2010b). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, 8, 70–91.
- Bennett, R. E. (2012). On the meaning of constructed response. In W. C. Ward & R. E. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Routledge.
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370–407. <https://doi.org/10.3102/0091732X14554179>
- Bennett, R. E. (2020). Interpreting test-score comparisons. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability issues in large-scale assessment* (pp. 227–235). National Academy of Education. <https://doi.org/10.31094/2020/1>
- Bennett, R. E. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice*, 43(4), 119–125. <https://doi.org/10.1111/emip.12652>
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 6(9), 3–38.
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer.
- Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Liggett, J., Marshall, G., & Zack, J. (1999). Using multi-media in large-scale computer-based testing programs. *Computers in Human Behavior*, 15, 283–294.

- Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007-466). National Center for Education Statistics, U.S. Department of Education. <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466>
- Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment*, 8(8), 3–44.
- Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). Routledge.
- Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment*, 25(3), 198–217. <https://doi.org/10.1080/10627197.2020.1804351>
- Bennett, R. E., Zhang, M., & Sinharay, S. (2021). How do educationally at-risk men and woman differ in their essay-writing processes? *Chinese/English Journal of Educational Measurement and Evaluation* | 教育测量与评估双语季刊, 2(1), 1–17. <https://www.ce-jeme.org/journal/vol2/iss1/1>
- Bennett, R. E., Zhang, M., Sinharay, S., Guo, H., & Deane, P. (2022). Are there distinctive profiles in essay-writing processes? *Educational Measurement: Issues and Practice*, 41(2), 55–69. <https://doi.org/10.1111/emip.12469>
- Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.). (2020). *Comparability issues in large-scale assessment*. National Academy of Education. <https://doi.org/10.31094/2020/1>
- Braun, H., Bejar, I. I., & Williamson, D. M. (2006). Rule-based methods for automated scoring: Application in a licensing context. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Lawrence Erlbaum Associates.
- Breithaupt, K. J., Mills, C. R., & Mellican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds.), *Computer based testing and the Internet* (pp. 219–251). Wiley.
- Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1), 27–40.
- Brown, T., Chen, J., Ali, U., Costanzo, K., Chung, S., & Ling, G. (2015). *Mode comparability study based on PARCC Spring 2014 field test data* [Unpublished manuscript]. ETS.
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation* | 教育测量与评估双语期刊, 5(3). DOI: <https://doi.org/10.59863/MIQL7785>

- Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Routledge.
- Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, 40(11), 4715–4729.
- Clauser, B. E., Margolis, M. J., & Clauser, J. C. (2016). Issues in simulation-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 49–76). Routledge.
- College Board. (2024, January 4). *Everything you need to know about the digital SAT*. <https://blog.collegeboard.org/everything-you-need-know-about-digital-sat>
- Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement*, 1(1), 59–64.
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
- Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*, 19–41.
- Csapo, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment (pp. 143–230). In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills*. Springer.
- Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (ETS Research Report No. RR-14-03). ETS. <http://dx.doi.org/10.1002/ets2.12002>
- Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018). *Behavioral differences between retyping, drafting, and editing: A writing process analysis* (ETS Research Memorandum No. RM-18-06). ETS.
- Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M. E., Bennett, R. E., Sabatini, J. P., & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing*, 32, 1575–1606.
- DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Council of Chief State School Officers.
- Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38.
- Dillon, G. F., & Clauser, B. E. (2009). Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). *Simulation in Healthcare*, 4(1), 30.
- Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions, pp. 1–33. <https://doi.org/10.1145/3561048>
- Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Springer-Verlag.

- Dorans, N. J. (2000). Scaling and equating. In H. Wainer (Ed.), *Computer adaptive testing* (2nd ed., pp. 135–158). Erlbaum.
- Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning score scales* (pp. 179–198). Springer.
- Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). American Council on Education/Praeger.
- Duque, M. (2017). *Is there a PARCC mode effect?* Center for Education Policy Research. <http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-parcc-mode-pdf>
- EdTech Strategies. (2015). *Pencils down: The shift to on-line & computer-based testing*. https://www.edtechstrategies.com/wp-content/uploads/2015/11/PencilsDownK-8_EdTech-StrategiesLLC.pdf
- Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Publishers.
- Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Routledge.
- ETS. (2021). *Best practices for constructed-response scoring*. https://www.ets.org/content/dam/ets-org/pdfs/about/cr_best_practices.pdf
- ETS, Pearson, & Measured Progress. (2016). *Final technical report for 2015 administration* (pp. 141, 150). PARCC. <https://www.state.nj.us/education/assessment/district/PARCCTechReport15.pdf>
- Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J., & Johanessen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97(2), 201–224.
- Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193–202.
- Fu, J., & Feng, Y. (2018). *A comparison of score aggregation methods for unidimensional tests in different dimensions* (ETS Research Report No. RR-18-01). ETS.
- Fu, J., Chung, S., & Wise, M. (2013). *Statistical report of Fall 2009 CBALTM writing tests* (ETS Research Memorandum No. 13-01). ETS.
- Fu, J., Wise, M., & Chung, S. (2012). *Statistical report of Fall 2009 CBALTM reading tests* (ETS Research Report No. RR-12-12). ETS.
- Futia, G., & Vetro, A. (2019). On the integration of knowledge graphs into deep learning models for more comprehensible AI—three challenges for future research. *Information* 2020, 11(2), 122. <https://doi.org/10.3390/info11020122>
- Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment*, 8(1), 27–41.
- Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *Edge*, 6(1), 3–19. <http://edgaps.org/gaps/wp-content/uploads/EDge-Light.pdf>

- Georgia Department of Education. (2018). *Georgia's application for the Innovative Assessment Demonstration Authority under section 1204 of the Elementary and Secondary Education Act (ESEA)*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Flexibility/Georgia_IADA_Application.pdf
- Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.
- Gobert, J. D., Moussavi, R., Li, H., Sao Pedro, M., & Dickler, R. (2018). Real-time scaffolding of students' online data interpretation during inquiry with Inq-ITS using educational data mining. In M. Auer, A. Azad, A. Edwards, & T. de Jong (Eds.), *Cyber-physical laboratories in engineering and science education*. Springer. https://doi.org/10.1007/978-3-319-76935-6_8
- Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, 91, 92–105.
- G20. (2019). *G20 AI principles, section 1, article 1.3. (Annex to G20 ministerial statement on trade and the digital economy)*. https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf
- Guo, H., Rios, J. A., Haberman, S., Liu, L. O., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, 29(3), 173–183.
- Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, 55, 194–216. <https://doi.org/epdf/10.1111/jedm.12172>
- Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, 44, 571–596. <https://doi.org/10.3102/1076998619856590>
- Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2020). Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining*, 12(1), 19–45. <https://doi.org/10.5281/zenodo.3911797>
- Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, 40, 254–273.
- Hao, J., & Mislevy, R. J. (2018). *The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design* (Research Report 18–28). ETS.
- Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43(2), 16–29. <https://doi.org/10.1111/emip.12602>
- Haskins, C. (2019). Gaggle knows everything about teens and kids in school. *Buzzfeed*. <https://www.buzzfeednews.com/article/carolinehaskins1/gaggle-school-surveillance-technology-education>

- Hattie, J., & Timperly, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112.
- Heffernan, N., & Heffernan, C. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4), 470–497.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (MET Project Research Paper). Bill and Melinda Gates Foundation. https://danielsongroup.org/sites/default/files/inline-files/Reliability_Observations_School_Personnel_Gates.pdf
- Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp.5–30). Springer.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger Publishers.
- Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 5(2), 3–49.
- Howard, R. A. (1960). *Dynamic programming and Markov processes*. Technology Press–Wiley.
- Iacus, S. M., King, G., & Porro, G. (2011). *Causal inference without balanced checking: Coarsened exact matching*. Political Analysis. <http://j.mp/iUUwyH>
- Institute for Education Sciences. (n.d.). *Progress in International Reading Literacy Study (PIRLS): Countries*. <https://nces.ed.gov/surveys/pirls/countries.asp>
- Institute for Education Sciences. (2018). *Writing assessment*. <https://nces.ed.gov/nationsreportcard/writing/>
- Irvine, S. H., & Kyllonen, P. C. (2010). *Item generation for test development*. Routledge.
- Jewsbury, P. A., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at Grades 4 and 8: Mode evaluation study*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf
- Johns, J., & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. *Proceedings of the National Conference on Artificial Intelligence*, 21, 163.
- Johnson, S. (2019). *Students in California log on in record numbers to take online state tests*. EdSource. <https://edsources.org/2019/students-in-california-log-on-in-record-numbers-to-take-online-state-tests/613321>
- Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science*, 38, 523–550.
- Kingston, N. M (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education*, 22(1), 22–37.

- Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, 57, 1813–1824.
- Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). Cambridge University Press.
- Kolen, M. J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6(2), 73–96. https://www.tandfonline.com/doi/abs/10.1207/S15326977EA0602_01
- Kolen, M. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). Springer.
- Kolen, M. J., & Brennan, R. L. (2004). *Testing, equating, scaling, and linking: Methods and practices*. Springer.
- Kuang, C. (2017). Can A.I. be taught to explain itself? *New York Times Magazine*. <https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html>
- LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika*, 83(1), 67–88.
- LaMar, M., Baker, R. S., & Greiff, S. (2017). Methods for assessing inquiry: Machine-learned and theoretical. *Design Recommendations for Intelligent Tutoring System*, 5, 137.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lee, Y., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessment in Education*, 2, 8. <https://doi.org/10.1186/s40536-014-0008-1>
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83–102.
- Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement Issues and Practice*, 13(1), 5–8. <https://doi.org/10.1111/j.1745-3992.1994.tb00778.x>
- Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode comparability study based on spring 2015 operational test data*. ETS.
- Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance* (pp. 139–183). Harper & Row.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.
- Lottridge, S. M., Nicewander, A. W., & Mitzel, H. C. (2010). Summary of online comparability studies for one state's end-of-course program. In P. Winter (2010), *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers (pp. 13–32). http://www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf.

- Madiega, T. (2019). *EU guidelines on ethics in artificial intelligence: Context and implementation* (PE 640.163). European Parliamentary Research Service. [https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI\(2019\)640163_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf)
- Maglieri, G., & Comande, G. (2017). Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *International Data Privacy Law*, 7, 243–265. <https://doi.org/10.1093/idpl/ix019>
- Margolis, M. J., & Clauser, B. J. (2006). A regression-based procedure for automated scoring of complex medical performance assessment. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–167). Erlbaum.
- Marianti, S., Fox, J.-P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6), 426–451. <https://doi.org/10.3102/1076998614559412>
- Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olsen-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Erlbaum.
- Milne, R. B., Kelly, S. A., & Webb, D. C. (n.d.). *Effect of adaptivity on learning outcomes in an online intervention for rational number tutoring, "Woot Math," for Grades 3–6: A multi-site randomized controlled trial*. Saga Education, Inc. <https://www.wootmath.com/efficacy>
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. ETS.
- Mislevy, R. J., & DiCerbo, K. E. (2012). Evidence centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (p. 13). Information Age.
- Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction*, 5, 253–282. <https://doi.org/10.1007/BF01126112>
- Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior*, 15, 335–374.
- Mislevy, R. J., & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement*, 49, 148–166.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2). <https://doi.org/10.1016/j.rmal.2023.100050>
- Mullis, I. V. S. (2019). *eTIMSS: The future of TIMSS*. Boston College. <http://timss2019.org/frameworks/framework-chapters/introduction/etimss-the-future-of-timss/>
- Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 1.1–1.29). IEA.

- National Center for Education Statistics. (2012). *Writing 2011: National Assessment of Educational Progress at Grades 8 and 12* (NCES 2021-470). U.S. Department of Education. <https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf>
- National Center for Education Statistics. (n.d.). *2014 Technology and Engineering Literary (TEL)*. U.S. Department of Education. https://www.nationsreportcard.gov/tel_2014/
- Northwest Evaluation Association. (2019). *NWEA announces through-year assessment, a first-of-its-kind solution that blends proficiency and growth measurement*. https://www.nwea.org/content/uploads/2019/10/NWEA_Through-Year_Assessment_PR_OCT2019.pdf
- Olson, L. (2019). *The new testing landscape: How state assessments are changing under the new federal Every Student Succeeds Act*. FutureEd. https://www.future-ed.org/wp-content/uploads/2019/09/REPORT_NewTestingLandscape-1.pdf
- Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multi-stage testing: Theory and applications* (pp. 371–390). Chapman & Hall/CRC.
- O'Reilly, T., Sabatini, J. P., & Wang, Z. (2019). Using scenario-based assessments to measure deep learning. In K. Millis, D. L. Long, J. P. Magliano, & K. Wiemer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp.197–208). Routledge.
- Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. <https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf>
- Organisation for Economic Co-operation and Development. (2018). *PISA 2015 test questions*. <https://www.oecd.org/pisa/test/pisa2015/#d.en.537240>.
- Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence, section 1, article IV, 1.3*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#_ga=2.100756866.2085259285.1559316172-1236900936.1559134188
- Padilla, J.-L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144.
- Paek, P. (2005). *Recent trends in comparability studies*. Pearson Educational Measurement.
- Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). *Effectiveness of Cognitive Tutor algebra I at scale* (WR-984-DEIES). Rand Corporation. <http://www.siia.net/visionk20/files/Effectiveness%20of%20Cognitive%20Tutor%20Algebra%20I%20at%20Scale.pdf>
- Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. *Frontiers in Artificial Intelligence and Applications*, 200 (1), 531–538. <https://doi.org/10.3233/978-1-60750-028-5-531>
- Pearson. (2017, January 10). *Final technical report for the 2016 administration*. Illinois State Board of Education.

- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
- Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30(3), 10–28.
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015, December). Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Vol. 1, pp. 505–513).
- Pommerich, M. (2007). Concordance: The good, the bad and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning score scales* (pp. 200–216). Springer.
- Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the “gold standard.” *Applied Measurement in Education*, 28(2), 130–142. <https://doi.org/10.1080/08957347.2014.1002920>
- Pressey, S. L. (1926). A simple apparatus which gives tests and scores—and teaches. *School and Society*, 23(586), 373–376.
- Pressey, S. L. (1927). A machine for automatic teaching of drill material. *School and Society*, 25(645), 549–552.
- Protalinski, E. (2020). *How Duolingo uses AI in every part of its app*. Venture Beat. <https://venturebeat.com/2020/08/18/how-duolingo-uses-ai-in-every-part-of-its-app/>
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using item response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, 35(1), 38–47.
- Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the *e-rater*® automated scoring engine and humans for demographically based groups in the GRE® General Test (Research Report 18-12). ETS. <https://doi.org/10.1002/ets2.12192>
- Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES groups. *Educational Measurement: Issues and Practice*, 31(4), 2–12.
- Rijmen, F. (2009). *Horizontal and vertical linking in a longitudinal design* (Research Memorandum 09-03). ETS.
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14, 249–255.
- Rivas, A., Fraile, J. M., Chamoso, P., González-Briones, A., Rodríguez, S., & Corchado, J. M. (2019). Students’ performance analysis based on machine learning techniques. In L. Uden, D. Liberona, G. Sanchez, & S. Rodríguez-González (Eds.), *Learning technology for education challenges* (LTEC 2019, Vol. 1011, pp.428–438). Springer.
- Rosenbaum, P. R., & Rubin, D. R. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.

- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
- Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, 17(1), 20–32.
- Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K–12 testing programs. *Educational Assessment*, 24(4), 286–304.
- Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers et al. *Practical Assessment, Research and Evaluation*, 9(1), 1–9. <http://pareonline.net/getvn.asp?v=9&n=1>
- Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research and Evaluation*, 9(1), 1–13. <http://pareonline.net/getvn.asp?v=9&n=10>
- Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (Eds.). (2005). *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-457). National Center for Education Statistics, U.S. Department of Education. <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457>
- Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerize adaptive testing: From inquiry to operation*. American Psychological Association.
- Sawyer, R. (2007). Some further thoughts on concordance. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning score scales* (pp. 217–230). Springer.
- Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of computer-adaptive and paper-and-pencil test scores on the GRE General Test* (Research Report 98-38). ETS.
- Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mill, C. N., & Durso, R. L. (1995). *The introduction and comparability of the GRE Computer-Adaptive General Test* (Research Report 95-20). ETS.
- Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1848–1858). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P16-1>
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215–232. <http://www.jstor.org/stable/1435044>
- Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data* (CRESST Report 807). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153–189.

- Shute, V. J., & Zapata-Rivera, D. (2010). Intelligent systems. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 75–80). Elsevier.
- Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, 34(1), 3–12.
- Sireci, S. G. (2020). Standardization and UNDERSTANDARDization in educational assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105.
- Sireci, S. G., & O’Riordan, M. (2020). Comparability issues in assessing individuals with disabilities. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability issues in large-scale assessment* (pp. 178–204). National Academy of Education. <https://doi.org/10.31094/2020/1>
- Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457–490.
- Skinner, B. F. (1958). Teaching machines. *Science*, 128(3330), 969–977.
- Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press.
- Smarter Balanced Assessment Consortium. (2023). *Usability, accessibility, and accommodations guidelines*. The Regents of the University of California.
- Smarter Balanced Assessment Consortium. (2024). *2024–2025 interim assessments overview*. The Regents of the University of California. <https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf>
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 151–166.
- Suppes, P. (1972). Computer-assisted instruction at Stanford [Conference proceedings]. In M. Marois (Ed.), *Man and computer* (pp. 298–330). Karger. https://suppes-corpus.stanford.edu/sites/g/files/sbiybj7316/f/computer-assisted_instruction_at_stanford_113.pdf
- Tucker, B. (2012). Grand test auto: The end of testing. *Washington Monthly*. http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php
- Turek, M. (n. d.). *Explainable AI (XAI)*. Defense Advanced Research Projects Agency. <https://www.darpa.mil/program/explainable-artificial-intelligence>
- United States Medical Licensing Examination. (2018). *Sample test questions: Step 3*. https://www.usmle.org/pdfs/step-3/Step3_Sample_Items.pdf
- U.S. Department of Education. (n.d.). *Race to the Top assessment program*. <https://www2.ed.gov/programs/racetothetop-assessment/index.html>
- U.S. Department of Education. (2012). *Writing 2011: National Assessment of Educational Progress at grades 8 and 12* (NCES-2012-470).
- U.S. Department of Education. (2018). *A state’s guide to the U.S. Department of Education’s assessment peer review process*.

- van der Linden, W. J. (2017). Test speededness and time limits. In *Handbook of item response theory* (pp. 249–266). Chapman and Hall/CRC.
- van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika*, 80(3), 689–706.
- Vandermeulen, N., Leijten, M., & Van Waes, L. (2020). Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, 12(1), 109–140. <https://doi.org/10.17239/jowr-2020.12.01.05>
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education*, 15(3), 147–204. <http://www.andestutor.org/Pages/AndesLessonsLearnedForWeb.pdf>
- VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education*, 8(2), 179–221.
- Van Meer, E. (2003). Plato: From computer-based education to corporate social responsibility. *Iterations: An Interdisciplinary Journal of Software History*, 2(1), 1–22. <http://www.cbi.umn.edu/iterations/vanmeer.pdf>
- Veldkamp, B. P., & Paap, M. C. S. (2017). Robust automated test assembly for testlet-based tests: An illustration with analytical reasoning items. *Frontiers in Education*, 7, 1–8. <https://www.frontiersin.org/articles/10.3389/feduc.2017.00063/full>
- von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, 44(6), 671–705.
- Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & Blas, C. A. W. (Eds.), *Computerized Adaptive Testing: Theory and Practice*, 245–269.
- Wainer, H., & Eignor, D. R. (2000). Caveats, pitfalls, and unintended consequences of implementing large scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 271–299). Erlbaum.
- Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78.
- Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38(1), 19–49.
- Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning*, 2, 271–282.
- Ward, W. C., Kline, R. G., & Flaugh, J. (1986). *College Board computerized placement tests: Validation of an adaptive test of basic skills* (Research Report 86-29). ETS.

- Way, D., & Strain-Seymour, E. (2021). A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress. American Institutes for Research. <https://www.air.org/sites/default/files/Framework-for-Considering-Device-and-Interface-Features-NAEP-NVS-Panel-March-2021.pdf>
- Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April 8–10). *Score comparability of online and paper administration of the Texas Assessment of Knowledge and Skills* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.
- Way, W. D., Lin, C., & Kong, J. (2008, March 25–27). *Maintaining score equivalence as tests transition online: Issues, approaches and trends* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.
- Way, W. D., & Robin, F. (2016). The history of computer-based testing. In C. S. Wells & M. F. Boyd (Eds.), *Educational measurement: From foundations to future* (pp. 185–207). The Guilford Press.
- Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April 10–12). *An evaluation of a matched samples method for assessing the equivalence of online and paper test performance* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States.
- Weiss, D. J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24–35). U.S. Civil Service Commission.
- Weiss, D. J. (1978). *Proceedings of the 1977 computerized adaptive testing conference*. University of Minnesota.
- Weiss, D. J. (1980). *Proceedings of the 1979 computerized adaptive testing conference*. University of Minnesota.
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.
- Weiss, D. J. (1985). *Proceedings of the 1982 computerized adaptive testing conference*. University of Minnesota.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.
- Winter, P. (2010). *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers. www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf
- Wise, L. L. (2011). Picking up the pieces: Aggregating results from through-course assessments. ETS. https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Wise.pdf
- Yeung, C. K., & Yeung, D. Y. (2018, June). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual*

- ACM Conference on Learning at Scale (pp. 1–10). Association for Computing Machinery.
- Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in performance between computer-based and handwritten essays* (RR-04-18). ETS.
- Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15(4), 337–362.
- Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice*, 38(2), 14–26. <https://doi.org/10.1111/emip.12249>
- Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report 15-27). ETS.
- Zhang, M., Deane, P., van Rijn, P. W., & Bennett, R. E. (2019). Scenario-based assessments in writing: An experimental study. *Educational Assessment*, 24, 73–90. <https://doi.org/10.1080/10627197.2018.1557515>
- Zhang, M., Zou, D., Wu, A. D., Deane, P., & Li, C. (2017). An investigation of writing processes employed in scenario-based assessment. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 321–339). Springer.

NOTES

1. Soft-max is an extension of the logistic function to multiple-category variables, for example, $p(x_i) = \frac{\exp x_i}{\sum_j x_j}$.
2. That USMLE uses both analysis of product and response process in computing scores raises the question of how automated scoring differs from response process analysis in the consequential testing context. As suggested in the section introduction, automated scoring is generally used to produce one or more quantities for input into computing a test score. Response process analysis, in contrast, has been more often directed at such purposes as providing validity evidence, identifying possible guessing or cheating behavior, describing how groups differ in their approaches to problem-solving, and suggesting how instruction might be redirected. Automated scoring is widely used in operational consequential assessment. Response process analysis is far less prominent.
3. The methods used here are like the machine learning methods described for the evaluation of response processes in the prior section.
4. Relatively few studies have looked at this issue, but those that have been conducted suggest the presence of substantively important differences in how algorithms operate across at least some demographic groups (Bridgeman et al., 2012; Ramineni & Williamson, 2018).

5. There are many examples of current assessments taking a more traditional mastery test form that are used to support instruction but are not of the type described here. Measures like the Smarter Balanced interim assessment blocks are linear tests constructed to measure mastery of a narrow competency (or set of related narrow competencies) (Smarter Balanced, 2019). Results are linked to instructional resources in the Smarter Balanced Tools for Teachers.
6. Note that the focus here is on learning interactions only. In contrast, some schools record virtually every student online interaction for purposes of identifying safety threats (Haskins, 2019).

Modeling Data From Educational Assessments

Li Cai

University of California, Los Angeles

Mark Hansen

University of California, Los Angeles

Minjeong Jeon

University of California, Los Angeles

Michael C. Edwards

Arizona State University

Data from educational assessments can take on a variety of forms and be collected for a variety of purposes. The models and modeling required to infer meaning from assessment data are also varied. To treat this topic in breadth and to give proper respect to the diversity and volume of past research in this area, we take the perspective that modelers of educational assessment data are generalist statisticians who are not driven by a particular preconceived model or epistemology of measurement (see Tal, 2020, and the extensive references therein) and are not married to a single philosophy about statistical modeling or theory of inference (as in Rubin, 1984). As empirical researchers, we remain flexible and pragmatic in the usage of our statistical tools, with the chief goals of advancing our collective understanding about educational processes and empowering the users of educational assessments with information that they need to pursue improvement in educational outcomes.

In many ways, this more expanded view on statistical modeling is a point of departure from what is arguably one of the closest conceptual relatives to this chapter in the fourth edition of *Educational Measurement*. Yen and Fitzpatrick's (2006) foundational piece goes into depth on many aspects of item response theory (IRT) that are useful for practitioners and researchers, but it is narrower in scope. We view IRT models as an extremely important subset of the series of models that we will consider in this chapter.

Given the development of large-scale predictive modeling arising out of data-intensive computer science (see Breiman, 2001, for a discussion of the two cultures of statistical modeling), recent critiques of quantitative research and the psychometric tradition (e.g., Dixon-Román & Gergen, 2013), and the juxtaposition of imperfect theories in the social sciences with simplifying assumptions often made in statistical modeling under resource constraints, we must stress that the word “model” should be understood in the information-theoretic sense—that is, a model is a mathematical encoding that binds together data and theory (see Bonifay & Cai, 2017; Fanelli, 2019; Grünwald, 2007). A modeler's success should be judged by the extent to which the model's parameters and outputs enable researchers to understand contexts, processes, and outcomes more clearly. While considering how well the model accounts for the observed data (model fit), one must also consider modeling principles such as parsimony and interpretability. Therefore, our technical approach is decidedly more traditional than the data-intensive algorithmic modeling that takes the form of statistical machine learning.

We begin the chapter with a broad discussion of the various sources and types of data derived from educational assessments to help illustrate the need for the statistical modeler to pay attention to the purpose of statistical inference, to represent salient features of research design in the model, and to guard against overcomplication. This is followed by a discussion of various mathematical and probabilistic tools that have found success in applied statistics, including in educational assessment. We then introduce a general framework for modeling. The fact that most general constructs covered by educational assessments are, by their very nature, not directly observable necessitates our reliance

on, and a review of, relevant latent variable modeling techniques. While we will not shy away from technical discussions, our general orientation is theory and design driven, as opposed to purely data driven.

As to the empirical examples, we will introduce and report our analyses of two recent assessment and evaluation studies. First, we take a deep dive into a data set collected from a randomized, controlled, multisite trial of a math intervention with pre and post measures. This study illustrates modeling issues in a realistically complex scenario that combines key elements of contemporary psychometric analysis (e.g., item analysis, dimensionality evaluation, measurement invariance) with those of educational statistics in support of sound research design (e.g., representations of initial status, growth, unit of treatment assignment, nested or crossed relationships). Second, in the context of a project that aims to develop a new assessment program for English learners with severe cognitive disabilities, we highlight the need for advanced statistical modeling in a highly constrained space, as defined by legal requirements and practical considerations. We conclude by suggesting possible trends that the reader should be aware of in the modeling of educational assessment data.

VARIED PURPOSES AND SOURCES OF DATA

As other chapters in this volume indicate, educational assessments with different purposes can yield distinct types of educational assessment data. In practice, one assessment program may serve multiple purposes, and those purposes may also change over time (Ho, 2014). Therefore, we do not intend to create an exhaustive or mutually exclusive list in this section. Our hope is that by enumerating data sources explicitly, we may point out different features of these data and consequently different methods of statistical analysis and modeling that may be required.

Different Sources

Educational assessment data could be generated by the individuals being assessed or by others (e.g., raters or computer algorithms) and concern the knowledge, proficiencies, practices, or processes that would be desirable (or sometimes mandatory) for the educator, education researcher, or policy maker to understand and, if applicable, to quantify. These data may involve different types of variables (from discrete to continuous), and various kinds of nesting or cross-classification relationships among units may be present in the data set.

Classroom Assessments

We highlight classroom assessments first because they exemplify the need for educational assessments to serve the goal of student learning (see Brookhart & DePascale, this volume). These assessments are ideally brief, initiated by and from instruction, and scored rapidly—with or without technology—to provide feedback

to the learning process. Their purposes may range from practicing to progress checking to the diagnosis of misunderstanding or misconceptions. The data coming from these assessment instruments are also typically not as standardized as some of the other sources we will discuss. However, with the increasing awareness of formative assessment and instructional practices—and with the proliferation of educational technology products that can aid the teacher's timely use and scoring of curriculum-informed assessments—classroom assessments can yield an important source of data that are among the most authentic indicators of student learning and achievement that the educational assessment researchers and users have not yet been able to fully harness. On the one hand, when the consequences of lack of comparability in the results are few and minor, the statistical treatment of classroom assessment data may only require simple statistical tools, such as a box plot and sample descriptive statistics. On the other hand, when serving research or policy purposes, the classroom assessment data may need to be aggregated and analyzed for comparative purposes. In such cases, more advanced treatment—perhaps using latent variable models that are based on the item response or diagnostic classification traditions—might be necessary.

Summative Assessments

In contrast with classroom assessments, summative assessments are almost always developed and administered under far more strict requirements of standardization and comparability. In the U.S. context, statewide summative assessments are federally mandated and are typically developed to measure a set of content standards that are adopted by a state and should be broadly implemented in the state's public schools, though the latter may not always be the case. The summative assessments aim to take a snapshot of the students' achievement toward the end of the school year against the backdrop of a set of achievement standards expected of the students. When aggregated up to the level of schools and districts and disaggregated by important student background characteristics, these assessment programs provide important accountability data for policy makers and for the public (see Ho & Polikoff, this volume). The modeling of summative assessment data typically requires modern psychometric models and statistical analysis approaches such as IRT to aid the validation process and to ensure technical quality of the assessment with respect to dimensionality, calibration, linking and equating, item and test bias, setting of standards, evaluation of classification accuracy, etc. Increasingly, summative assessments are delivered with educational technology. Computerized adaptive administration methods have become increasingly common, requiring more advanced analytic techniques for development and maintenance (see Shermis et al., this volume). The particular psychometric approaches taken by the analyst may also depend on whether the assessment (or parts of the assessment) is being developed and field tested or is being maintained and evaluated on an ongoing operational basis.

Large-Scale Sampled Surveys

Large-scale sampled education surveys play important roles in education policy and research. National and international assessment programs that resemble the National Assessment of Educational Progress (NAEP) have proliferated since at least the beginning of the 21st century (see Ho & Polikoff, this volume). Other longitudinal data collection programs, such as the Early Childhood Longitudinal Study, have provided major sources of data for researchers and policy analysts. They combine features of typical educational assessments delivered under standardized conditions, with scientific and representative sampling of students, schools, and education agencies. Unlike standardized assessments for summative evaluation purposes, NAEP-like assessments often include measures of both student achievement and development, along with a host of background measures that can help policy makers and researchers understand the relationship and interactions among input variables and student outcomes (see also Braun & Kirsch, this volume, for a discussion of international large-scale survey assessments). To the individual student participants, these large-scale sampled surveys are of low stakes. But because they help provide broad gauges of the performance of education systems, their design and analysis often involve highly complex combinations of IRT modeling, latent regression analysis, and design-based inference (e.g., complex survey design adjustments).

Evaluation Studies

Evaluation studies, including randomized controlled trials and quasi-experimental evaluation studies of the effectiveness of programs and interventions, have become routine in education sciences. The emergence of notable outlets such as the *Journal of Research on Educational Effectiveness* is a testament to the change in landscape. Researcher-developed assessments are often used to measure outcomes of an intervention, as well as important correlates, mediators, and moderators. For instance, a randomized multisite trial of a math intervention might involve an assessment of student learning in the targeted domain(s) as the outcome(s) of interest. The fidelity of program implementation might be measured at the site (school) level to help understand treatment effect variation, using observational rubrics and protocols that result in quantifiable assessment items. The teachers' professional development related to the intervention may also be measured using survey items, to help understand their mediating effects. These assessment tools are routinely validated using a combination of classical and modern psychometric techniques, and the resulting scores may be modeled separately using multilevel/hierarchical regression. More recently, joint modeling and estimation methods have been developed that combine the data from these assessments with important design characteristics (e.g., treatment assignment indicator) to yield more informative inference about intervention effect (see Cai, Choi, et al., 2016).

Technology-Based Assessments

Since the beginning of the 21st century, interactive educational technology has been increasingly involved in the design and delivery of educational assessments. This technology not only offers researchers with a far richer set of templates to conceptualize what constitutes an assessment item, including game-like environments or simulations, but also brings the potential of a deeper coupling of assessment with instruction that can be scaled up with a high degree of quality not achievable via traditional means. Importantly, technology allows for the collection of what is known as *process data*, through embedding software code into the assessment delivery platform itself (see Bennett et al. and Ercikan & Solano-Flores, this volume). The process data provide insights well beyond the traditional right/wrong answers; they reveal the process of problem solution and can suggest feedback to the learner or the teacher that has greater diagnostic value. They also tend to be intensively longitudinal in nature. A relatively simple kind of process data, response time, has already received a considerable amount of attention in the literature. Models for other types of process data have also been studied, with general latent variable models with sequential dependencies proving particularly interesting in this context (see Bergner & von Davier, 2019).

Different Purposes and Modeling Choices

The eminent statistician John W. Tukey once said, “The best thing about being a statistician is that you get to play in everyone’s backyard” (Brillinger, 2014). This one-liner epitomizes the role statistical modelers can play in research—statistical modeling supports the full educational assessment development life cycle. We believe that the thorough participation of and input from statistical modelers in assessment design, validation, and use leads to better results than treating statistical analysis as merely one isolated step in the process. Conversely, we note that competent statistical modeling in support of educational assessment development requires a broad understanding of research design and its substantive objectives. Some tasks may require more complex models, but some—either because of resource or timing constraints—may require simpler alternatives. Ideally, the analyst should be equipped with an adequately general and flexible framework—one that can also be constrained and specialized to accommodate the specific situation at hand.

To this end, we provide Table 10.1 to help guide the reader, in a similar manner as Bartholomew et al. (2011) did in their introductory chapter or akin to Thissen and Steinberg’s (1986) taxonomy. In Table 10.1, we attempt to synthesize a number of widely used and “named” models in educational assessment research, as well as ones that have not yet been named. For example, on the one hand, with a single-level data collection design, all observed variables being categorical in nature, and the assumed underlying latent variable(s) being continuous, an appropriate model may

Table 10.1 An Overview of More Specialized Models in the General Framework

Latent Variables	Predictors Present	Data Structure							
		Single Level		Nested Multilevel		Cross-Classified		Observed Variables	
		Observed Variables		Observed Variables					
		Categorical	Continuous	Categorical	Continuous	Categorical	Continuous		
Continuous only	No	IRT model	Factor model	Multilevel IRT / hierarchical / two-tier item factor model	Multilevel factor model / bifactor model	Cross-classified IRT model	G theory model		
	Yes	Latent regression model	MIMIC model	Generalized and nonlinear random effects model	Hierarchical linear model / linear mixed model	Explanatory IRT model	Cross-classified linear mixed model		
Categorical only	No	DCM	Latent class model	Multilevel DCM	Multilevel latent class model				
	Yes								
Categorical and continuous	No	Mixture item factor model	Factor mixture model	Multilevel mixture IRT model	Multilevel factor mixture model				
	Yes								

Note. DCM = diagnostic classification model; G theory = generalizability theory; IRT = item response theory; MIMIC = multiple indicators and multiple causes. Gray cells = generalized latent variable models.

be an IRT model. On the other hand, if the design implies a cross-classified data structure (e.g., items are regarded as random samples and predictors are present for one or more of the random item effects in the model), an explanatory IRT model (de Boeck & Wilson, 2004) may be applicable. The reader should note that large swaths of the table are grayed out. This fact merely implies that while appropriate and potentially useful generalized latent variable models may be specified, they are not yet so widely used.

CORE STATISTICAL MODELING CONCEPTS

In this section, we introduce some core concepts that are fundamental to the modeling of complex multivariate data that are characteristic of educational measurement. These include latent variables, conditional independence, likelihood, prior, posterior, information, and model fit evaluation. It is helpful to keep in mind that we discuss these concepts to equip us with the knowledge to accomplish the major tasks of modeling data in educational assessment:

- model building, which is about the choices that one makes to create a model to reflect the design and the desired inference from the assessment data;
- estimation and inference, which is about systematically optimizing the parameters in a proposed model, given the data at hand and the choice of optimality; and
- quantification of error and uncertainty, which deals with the tolerance for error and uncertainty with respect to inference (primarily) and model fit (to the extent that the quality of inference is affected).

We assume that the reader has some knowledge in elementary probability and statistics and will elaborate on technical developments when possible.

The Ubiquity of Latent Variables

A latent variable is a random variable that is not observed but is, nevertheless, specified in a statistical model. As part of model formulation, latent variables are often posited in the analysis of educational assessment data because the constructs involved (e.g., proficiency in math or reading) are not as discernable as markers such as height or the amount of cholesterol per unit of blood. Latent variable models may include either continuous or discrete random variables or a mixture of both kinds in one model. The use of latent variables is so pervasive that statisticians often overlook their many incarnations. However superficially dissimilar, latent variables may appear as common factors in factor analysis and structural equation modeling, latent traits in IRT (which are also common factors), latent attributes in diagnostic classification models, incomplete/missing data in both survey nonresponse and statistical computation, and random effects in multilevel modeling. In all these cases, a distribution of the latent variable is assumed. This could be one of the “named” distributions

(e.g., normal or multivariate normal) or a discrete distribution with point probability masses on particular values (e.g., 0–1). In some cases, the distribution can be semiparametric or seminonparametric in nature (e.g., Woods & Thissen, 2006). We would like to mention that the latent variables we introduce in this chapter follow the *reflective indicators* mindset prevalent in psychometrics. For a recent discussion on reflective versus formative indicators, the reader is referred to Bollen and Diamantopoulos (2017).

The latent variables are connected to the observations through a conditional distribution. This conditional distribution specifies the probability of the observed assessment data, given a particular value of the latent variables. In educational measurement, this conditional distribution can be understood as the measurement model. Technically, one can view the unknown parameters in the measurement model as latent variables as well. This is especially helpful in a fully Bayesian approach to estimation and inference, such as in data-augmented Markov chain Monte Carlo algorithms (e.g., Albert, 1992). Most of the time, it is helpful to keep the distinction between parameters of a structural nature and those of an incidental nature (Neyman & Scott, 1948) and to index these separately. In the next few sections, several examples are used to help illustrate the structural/incidental distinction. Specifically, characteristics of assessment items are regarded as structural, but individual-specific differences (e.g., standing of test takers with respect to the latent variable or variables) are treated as incidental. We also discuss the idea of measurement models as conditional probability distributions on latent variables.

Factor Analysis as Regression With Unobserved Predictors

Our first example is a classical one. In both Thurstone's (1935) common factor model and subsequent developments of confirmatory factor analysis and structural equation modeling (Jöreskog, 1969), observed variables are assumed to be a linear combination of continuous latent variables. Let y_i denote a generic (continuous) observed variable i and let us consider the case of p common factors indexed $\theta_1, \dots, \theta_p$; Equation 1 indicates that y_i depends on the common factors θ and the unique factor e_i ,

$$y_i = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \dots + \lambda_{ip}\theta_p + e_i. \quad (1)$$

The weights λ of the linear combination are known as factor loadings, already familiar to education researchers. We consider them to be structural parameters. The common factors are the latent predictors in this regression model. Typically, the θ 's are assumed to have zero mean and unit variance, and the unique factor variance (say σ_i^2) is determined by the model as a residual variance component, representing variance in the outcome variable that is not accounted for by the predictors. Just as in linear regression modeling, the unique factor is often assumed to have a normal distribution with mean zero, effectively

becoming a disturbance term. This should make it immediately clear that the conditional distribution of y_i given values of θ is univariate normal, that is,

$$y_i | \theta \sim N(\lambda_i' \theta, \sigma_i^2), \quad (2)$$

where $\theta = (\theta_1, \dots, \theta_p)$ is a $1 \times p$ vector of random variables, $\lambda_i' = (\lambda_{i1}, \dots, \lambda_{ip})$ is a $p \times 1$ vector of factor-loading parameters, and $\lambda_i' \theta = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \dots + \lambda_{ip}\theta_p$ is the vector “inner” product between λ_i and θ .

Item Factor Analysis and Multidimensional IRT

The second example builds on the linear model in Equation 1 and retells the story contained in Bock et al. (1988), Takane and de Leeuw (1987), and Bock and Aitkin (1981). In this example, the observed variable is discrete or categorical instead of continuous. To begin, let us consider the following regression model with the underlying continuous unobserved response process variate y_i^* as the outcome and θ 's as the predictors. As before, we assume that the θ 's are a set of uncorrelated standard normal variates, and the regression equation is

$$y_i^* = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \dots + \lambda_{ip}\theta_p + e_i^* = \lambda_i' \theta + e_i^*. \quad (3)$$

For convenience, we shall let the error term e_i^* be normally distributed with mean zero and unique variance $\sigma_i^2 = 1 - \lambda_i' \lambda_i$. As is customary, we also assume that e_i^* is uncorrelated with the θ 's, such that y_i^* is conveniently scaled to have unit variance. The foregoing assumptions imply that conditional on θ , y_i^* is normally distributed with conditional variance $\sigma_i^2 = 1 - \lambda_i' \lambda_i$ and conditional mean $\lambda_i' \theta$.

Let us further assume that there is a threshold τ_i , and that the observed response y_i becomes positive/correct/endorsed ($y_i = 1$) if and only if $y_i^* > \tau_i$. We now see that the probability of a response of 1 from a subject with latent variable θ is equal to the following normal integral:

$$P(y_i = 1 | \theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \int_{\tau_i}^{\infty} \exp\left(-\frac{(y_i^* - \lambda_i' \theta)^2}{2\sigma_i^2}\right) dy_i^*. \quad (4)$$

We can now make a change of variable for the integral by defining $z = (y_i^* - \lambda_i' \theta) / \sigma_i$, and that leads us to denote the probability of a response of 1 as

$$P(y_i = 1 | \theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\eta_i} \exp\left(-\frac{z^2}{2}\right) dz, \quad (5)$$

where

$$\eta_i = \frac{-\tau_i + \lambda_i' \theta}{\sigma_i}$$

is often known as the *linear predictor* in the generalized linear modeling literature (McCullagh & Nelder, 1989). But the right-hand side of Equation 5 is simply the standard normal cumulative distribution function with argument η_i , that is,

$$P(y_i = 1 | \boldsymbol{\theta}) = \Phi(\eta_i) = \Phi(\alpha_i + \boldsymbol{\beta}_i' \boldsymbol{\theta}),$$

for unstandardized item location or intercept parameter

$$\alpha_i = -\frac{\tau_i}{\sigma_i},$$

and unstandardized vector of item discrimination or slope parameters

$\boldsymbol{\beta}_i = \sigma_i^{-1} \boldsymbol{\lambda}_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{ip})$, with $\beta_{ip} = \frac{\lambda_{ip}}{\sigma_i}$. Using terminology from generalized linear models (McCullagh & Nelder, 1989), we would say that the model in Equation 5 uses a probit link function. Jöreskog and Moustaki (2001) provided an extensive discussion of this underlying normal approach, as well as the associated estimation methods.

We can approximate the model in Equation 5 with Birnbaum's (1968) logistic substitution, leading to a multidimensional IRT (MIRT) model that is widely familiar in educational measurement (Reckase, 2009),

$$P(y_i = 1 | \boldsymbol{\theta}) = \frac{1}{1 + \exp[-1.702(\alpha_i + \boldsymbol{\beta}_i' \boldsymbol{\theta})]}, \quad (6)$$

where the constant of 1.702 rescales the logistic parameters approximately to the probit/normal metric (see Camilli, 1994, for a recount of the history behind the approximation). Recall that we have assumed the $\boldsymbol{\theta}$'s to be uncorrelated. If we choose to estimate the more natural unstandardized item location and slope parameters, the standardized parameters in threshold and loading form can be recovered by these simple transformations (see Wirth & Edwards, 2007 or Cai, 2010a):

$$\tau_i = -\frac{\alpha_i}{\sqrt{1 + \boldsymbol{\beta}_i' \boldsymbol{\beta}_i}}, \quad \lambda_{ip} = \frac{\beta_{ip}}{\sqrt{1 + \boldsymbol{\beta}_i' \boldsymbol{\beta}_i}}.$$

If there is only a single latent variable, Equation 6 reduces to the familiar two-parameter logistic (2PL) model:

$$P(y_i = 1 | \theta) = \frac{1}{1 + \exp[-(\alpha_i + \beta_i \theta)]}. \quad (7)$$

In this case, the classical item difficulty parameter is equal to $-\alpha_i / \beta_i$, which corresponds to the value of θ at which the probability of the response of 1 is exactly 0.5.

Notice that as we complete the transition from the logistic as an approximation to the normal (from Equation 5 to Equation 6), we dropped the scaling factor of 1.702. The logistic model and its parameters form an entirely appropriate metric on which one can express item analysis results, and the scaling factor can and generally should be omitted in new educational assessment development projects to avoid misunderstanding or error. Obvious exceptions would include cases such as the maintenance of legacy banks of items calibrated under the normal ogive model.

Two-Tier and Item Bifactor Analysis

We now see that MIRT and item factor models are indeed deeply connected. In a certain sense, the distinction between MIRT and factor analysis is an artificial one largely resulting from historically separate software implementations. Realizing the underlying connection is of practical importance. In the development of educational assessments, it is imperative that the structural properties of the measurement instruments be examined. Usually this examination is treated as a part of dimensionality analysis, separate from considerations relating to item analysis, scaling, and test scoring. Unidimensional IRT modeling has become the default option for the latter, while factor analytic models dominate in dimensionality analysis. However, recent advances in item factor analytic methods and statistical computation have enabled the simultaneous modeling of item properties within a multidimensional measurement model.

As an example, consider Cai's (2010a) two-tier item factor model, which may be understood as a superset of the popular item bifactor model (Cai et al., 2011; Gibbons & Hedeker, 1992; Reise, 2012). In this model, the β_i 's form a particular pattern. Each item may load on two kinds of factors: the primary (Tier 1) factors, of which there can be any number, and a single group/specific (Tier 2) factor. When organized into a matrix, the item slopes may appear as the following for a six-item test:

$$\left(\begin{array}{cc|cc} \beta_{11} & & \beta_{13} & \\ \beta_{21} & & \beta_{23} & \\ \beta_{31} & & & \beta_{34} \\ \beta_{41} & \beta_{42} & & \beta_{44} \\ & \beta_{52} & & \beta_{55} \\ & \beta_{62} & & \beta_{65} \end{array} \right). \quad (8)$$

To the left of the vertical line, the primary factors form a familiar independent cluster pattern, perhaps representing two major domains or subjects, save for the fourth item, which cross-loads on both primary factors. To the right, there are three group/specific factors that influence three distinct item doublets, representing additional shared variation above and beyond the primary factors. A bifactor model has a single

primary factor. Wainer et al.'s (2007) testlet response theory model is a constrained version of the bifactor model, with equality (or proportionality) constraints on the slope parameters within each item and freely estimated variance components for each group/specific factor.

The parameters of item factor models such as those above may be estimated directly from educational assessment data, such as when field testing new items. When assessment developers are equipped with the knowledge that these methods exist and can support operational educational assessments at scale, new assessment programs may benefit directly by selecting models that appropriately reflect and correspond to the assessment design. The results of item factor analysis inform both item analysis and dimensionality analysis.

Generalized Partial Credit and Three-Parameter Logistic Models

Some educational assessments involve multiple categorical response data, rather than (or in addition to) dichotomously scored items. Examples include ordinal ratings of constructed responses, such as essays or short answers. Polytomous IRT models such as the generalized partial credit (GPC) model (Muraki, 1992) may be used in these cases. Suppose item i has C_i categories, with observed response $c \in \{0, \dots, C_i - 1\}$. A version of the GPC model may be written as

$$P(y_i = c | \theta) = \frac{\exp(c\beta_i\theta + \alpha_{ic})}{\sum_{m=0}^{C_i-1} \exp(m\beta_i\theta + \alpha_{im})}, \quad (9)$$

where the item slope β_i serves the same role as the discrimination parameter of a 2PL model. Since the category-specific intercept α_{ic} parameters are not identified without further constraints, the first intercept is typically fixed to 0 ($\alpha_{i0} = 0$). More broadly, the GPC model is one of many models in the nominal categories model family (Thissen & Steinberg, 1986). For further details, the reader is referred to Thissen et al.'s (2010) exposition.

For multiple-choice items that are dichotomously scored, the three-parameter logistic model (3PL; Birnbaum, 1968) can be used to account for the effect of guessing. The parameter γ_i is introduced to serve as a lower asymptote of the traceline:

$$P(y_i = 1 | \theta) = \gamma_i + \frac{1 - \gamma_i}{1 + \exp[-(\alpha_i + \beta_i\theta)]}. \quad (10)$$

When the lower asymptote is zero, Equation 10 becomes the familiar 2PL IRT model.

For an item with multiple categorical responses $c \in \{0, \dots, C_i - 1\}$, the conditional probability of observing response y_i is a multinomial with trial size 1 in C_i categories, that is,

$$f(y_i | \theta) = \prod_{c=0}^{C_i-1} P(y_i = c | \theta)^{x_c(y_i)}, \quad (11)$$

where the indicator function $\chi_c(y_i)$ takes a value of 1 if and only if $y_i = c$ and it is zero otherwise. This result parallels Equation 2.

Diagnostic Classification Models

The last example is more contemporary. There is recent attention on diagnostic classification models (DCMs), and book-length volumes have appeared in the literature (e.g., Rupp et al., 2010; von Davier & Lee, 2019). It is useful to review these models here because they highlight the possibilities afforded by explicitly constraining parameters in a measurement model, another important concept in modeling of data from educational assessments. DCM applications are more often confirmatory in nature than in item factor analysis, for instance. In contrast to the IRT model discussed earlier, where θ 's are continuous latent variables, standard DCMs assume the presence of discrete latent variables underlying the responses to items, often 0–1. They may represent mastery of skills or knowledge, presence of attributes, possession of misconceptions, etc.

Consider a $p \times 1$ vector of dichotomous (0–1) latent variables ξ . The relationship between items and latent attributes is captured by the so-called **Q**-matrix (Tatsuoka, 1983), which is a matrix of fixed 0's and 1's, just as a design matrix. The (i, q) th entry in the **Q**-matrix takes on a value of 1 if item i measures latent variable q . These values are typically informed by the blueprint and item development process. From the outset, a DCM resembles a MIRT model, which in turn extends a generalized linear model:

$$P(y_i = 1 | \xi) = \frac{1}{1 + \exp[-\eta_i]} = \frac{1}{1 + \exp[-\alpha_i - \beta_i' h(\mathbf{Q}, \xi)]}. \quad (12)$$

The $h(\mathbf{Q}, \xi)$ term in Equation 12 is a vector-valued function that defines how the latent variables in ξ should combine to create the linear predictor η_i . Henson et al. (2009) systematically placed constraints on the item parameters α_i and β_i and showed how different choices of $h(\mathbf{Q}, \xi)$ lead to a general model that contains several of the more commonly utilized diagnostic models as special cases (see also Choi et al., 2013; Rupp et al., 2010).

For instance, suppose that according to the **Q**-matrix, a mathematics test item measures two skills (attributes ξ_1 and ξ_2) and that successful solution of the item requires both skills. The linear predictor may take the well-known deterministic-input noisy “and” gate (DINA; Junker & Sijtsma, 2001) form, with an intercept and a single interaction term:

$$\eta_i = \alpha_i + \beta_i \xi_1 \xi_2. \quad (13)$$

In this case, $h(\mathbf{Q}, \xi) = \xi_1 \xi_2$ yields a second-order interaction only. Alternatively, perhaps the item only requires *either* Attribute 1 or Attribute 2. Then the deterministic-input noisy “or” gate (DINO; Templin & Henson, 2006) model may be appropriate,

with the slopes of the two main effects and the interaction constrained equal in absolute magnitude, albeit with the interaction having an opposite sign:

$$\eta_i = \alpha_i + \beta_i \xi_1 + \beta_i \xi_2 - \beta_i \xi_1 \xi_2. \quad (14)$$

One can easily verify that in terms of the log-odds, Equation 14 is indifferent as to which of the two attributes is present, and having both present provides no advantage in log-odds. Yet another possibility is reflected by the general diagnostic model (Hartz, 2002; von Davier, 2005), which is similar to factor analysis:

$$\eta_i = \alpha_i + \beta_{i1} \xi_1 + \beta_{i2} \xi_2. \quad (15)$$

The values of the latent attributes can be further modeled along the lines of higher order factor analysis (see, e.g., de la Torre & Douglas, 2004). The attributes essentially become hidden items, whose pattern of presence or absence can help identify their location relative to a second-order latent dimension (Cai, Choi, & Kuhfeld, 2016). For example, if there are $l=1, \dots, L$ attributes, the following one-parameter logistic (Rasch-type) model and a single higher order dimension ζ may be specified:

$$P(\xi_l = 1 | \zeta) = \frac{1}{1 + \exp[-(\tau_l + \zeta)]},$$

where τ_l is the intercept/location parameter for attribute l . Alternatively, the 2^L attribute profile probabilities may be represented as a multinomial with $2^L - 1$ free parameters. The economy in using the higher order formulation is clear; the model above accounts for the attribute profile with only $L + 1$ parameters. DCMs can also be further integrated with item factor analysis models to provide even more flexible modeling frameworks. These will be discussed further later in the chapter.

Conditional Independence

It is obvious that without knowing the value of the latent variable predictors in any of the models we have discussed, the measurement models are not directly estimable. After all, we only have access to the information contained in the observed variable, and all the previously presented equations contain more unknowns than available data. However, an aspect that we have not yet discussed is that most data in educational measurement are multivariate in nature, with assessments usually containing many items or tasks and each response possibly scored for different outcomes. Furthermore, the random variables in the multivariate data set usually present themselves to the statistician as correlated and patterned. This stands in contrast to, say, a traditional linear model, where the outcome variable y has a trivial covariance/correlation structure, or even a traditional hierarchical linear model (Raudenbush & Bryk, 2002), which may have several more variance or covariance components but retains the emphasis on modeling changes in expectations, as in regression analysis. The task of the statistician, therefore,

lies in the formulation of linear, generalized linear, or nonlinear multivariate statistical models that can represent the pattern of correlatedness in a meaningful way in terms of the underlying latent variables, ideally according to assessment design. When combined with the assumed distributional form of the latent variables, a joint distribution of the observed variables and latent variables naturally leads to the marginal distribution of the observed variables, which contains a host of information about the structural parameters. Finally, with Bayes' formula, we can find posterior distributions of the latent variables given the observed data and the model. Statistical summaries of the posterior, such as means and standard deviations, have important uses in the scoring and inference of educational assessments.

To accomplish the model building and statistical computation necessary in our kind of multivariate analysis for modeling educational assessment data, the most useful device at our disposal is arguably conditional independence, also known as residual independence and local independence in more restricted contexts. We illustrate it with two examples.

Exploratory (Unrestricted) Factor Analysis: Uncorrelated Residuals

First, recall that the linear factor analysis model in Equation 1 implies a conditionally normally distributed y , for a generic outcome variable (Equation 2). Let us assume that there exist n such variables, indexed as $y_1, \dots, y_i, \dots, y_n$, perhaps representing n measures. The vector of observations $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)$ can be represented in a multivariate regression model such as

$$\mathbf{y} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1p} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{np} \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \mathbf{\Lambda} \boldsymbol{\theta} + \mathbf{e},$$

where $\mathbf{\Lambda}$ is the $n \times p$ factor loading matrix, and $\mathbf{e} = (e_1, \dots, e_i, \dots, e_n)$ is the vector of unique factors, typically assumed to be multivariate normal with a zero mean vector and covariance matrix $\mathbf{\Psi}$, if, for example, maximum likelihood estimation and inference is desired. Consequently, the conditional distribution of $\mathbf{y}$ is multivariate normal:

$$\mathbf{y} | \boldsymbol{\theta} \sim \text{MVN}(\mathbf{\Lambda} \boldsymbol{\theta}, \mathbf{\Psi}). \quad (16)$$

We now introduce the crucial conditional independence assumption for exploratory factor analysis, which states that conditionally on $\boldsymbol{\theta}$, the variates in $\mathbf{y}$ become independent. In other words, if we could partial out/control for the influence of the latent variables, the observed variables should not be related at all. This affords a particular causal interpretation, that is, the observed associations are due *solely* to the influence of the hidden variables. Because of the normality assumption, uncorrelatedness is equivalent to independence. Thus, conditional independence in the

present context effectively becomes an assumption on the covariance structure of the unique factors: that $\mathbf{e}$ has a diagonal covariance matrix, that is, $\mathbf{\Psi} = \text{diag}(\sigma_i^2)$, with zero covariances.

The key insight is this: When the number of factors p is substantially fewer than the number of observed variables n , the factor analysis model equipped with conditional independence provides a much more parsimonious representation of the observed pattern of covariances/correlations, and the factor loadings can be identified from the observed covariance matrix, even though at first glance the exploratory factor model contains, paradoxically, more unknowns than there are available data. Together with other criteria such as Thurstone's simple structure and factorial invariance, the loading matrix provides a more meaningful explanation of the empirical data than the raw correlations alone.

Local Independence in IRT

The second example is also familiar to educational measurement researchers and helps set the stage for a discussion of likelihood, prior, marginal likelihood, and posterior. We begin with the 2PL-inspired MIRT model in Equation 6. For a single dimension θ , the model for a particular item i can be alternatively written, without loss of generality, as

$$P(y_i = 1 | \theta, \omega) = \frac{1}{1 + \exp[-(\alpha_i + \beta_i \theta)]} = T_i(\theta, \omega), \quad (17)$$

where the T_i notation pays homage to the term *traceline* coined by Lazarsfeld (1950). The parameters for this item are assumed to come from an overall parameter vector ω , to avoid notational clutter. This leads to a conditional Bernoulli distribution:

$$P(y_i | \theta, \omega) = [T_i(\theta, \omega)]^{y_i} [1 - T_i(\theta, \omega)]^{1 - y_i}, \quad (18)$$

where $y_i \in \{0, 1\}$, for example, an observed item response for a dichotomously scored test.

With n items, the local independence assumption in IRT states that for θ values sufficiently close to one another, that is, controlling for the latent variable, the collection of item responses should be independent. The IRT model, just as in factor analysis, tells a specific causal story about why the observed item responses are correlated. We realize that with conditional independence on θ , we can now write the conditional probability of the item response vector $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)$ in a fully factored form,

$$P(\mathbf{y} | \theta, \omega) = \prod_{i=1}^n [T_i(\theta, \omega)]^{y_i} [1 - T_i(\theta, \omega)]^{1 - y_i}, \quad (19)$$

which functions in the same way as the assumption of uncorrelated residuals in exploratory factor analysis that reduces the unique factor covariance matrix Ψ to a diagonal one, though without the help of multivariate normality that equates uncorrelatedness with independence. Evidence that suggests the lack of conditional or local independence (e.g., W. H. Chen & Thissen, 1997) causes concern in IRT applications because the resulting parameter estimates, as well as inferences based on those estimates, may be biased. In a similar vein, if the specification of the number of dimensions is incorrect in a factor analysis, the factor loadings will be biased. Thus, a large part of fit appraisal for latent variable models focuses on dimensionality specification.

Prior and Likelihood

In Equations 16 and 19, the conditional distributions of the observed variables on latent variables are specified. The model is still incomplete because the conditional distribution alone does not allow us to estimate the parameters and conduct inference about the latent variables. We must specify the prior distributions of the latent variables. In some texts, especially in IRT, they are also referred to as the population distributions. Regardless, the priors enable us to write down the joint probability of both observed and latent variables. Integrating over the latent variables leads to the marginal likelihood, which is free of latent variables.

Normal-Normal Factor Analysis Model

As an example, consider an exploratory factor analysis (Spearman, 1904). The initial factor extraction almost always assumes uncorrelated common factors; oblique factors are the result of rotation only. Thus, we can assume θ to be multivariate normal with zero means and an identity matrix as its covariance matrix. Combined with the conditional distribution in Equation 12, we arrive at the important marginal distribution of $\mathbf{y}$, free of any latent variables:

$$\mathbf{y} \sim \text{MVN}(\mathbf{0}, \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}), \quad (20)$$

which says that the covariance matrix of the observed variables is structured by the parameters of the model (its loadings and unique variances). The marginal distribution of $\mathbf{y}$ enables one to write down the marginal likelihood of the observed data, forming the basis of parameter estimation and model fit appraisal. Suppose we use Σ to represent the covariance matrix of $\mathbf{y}$. Per Equation 20, we have $\Sigma(\omega) = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$. In this case, we are using the parentheses to emphasize that in a similar manner as in Equation 13, we have a covariance structure model that depends on a set of structural parameters in ω , including the factor loadings and unique factor variances.

As an aside, to obtain Equation 20 we have taken advantage of our more modern knowledge of matrix algebra and multivariate normal distribution theory, which facilitates the derivation of the joint distribution of both $\mathbf{y}$ and θ . With joint normality, a

traditionally difficult task of integrating over the latent variable to arrive at the marginal distribution of $\mathbf{y}$ in Equation 20 can be accomplished with ease. This is not true in general.

Logistic-Normal IRT Model

The IRT model offers another, perhaps more direct, example. Continuing from Equation 19, if we assume the prior/population distribution of the latent variable θ is $h(\theta|\omega)$, typically normal with mean zero and unit variance (but can contain unknown structural parameters, if the data collection design requires their presence), the marginal probability of the observed item response vector is

$$P(\mathbf{y}|\omega) = \int P(\mathbf{y}|\theta, \omega) h(\theta|\omega) d\theta. \quad (21)$$

By a perversion of notation involving data and parameters, we see that Equation 21 provides the marginal likelihood of the item parameters ω , given observed data $\mathbf{y}$:

$$L(\omega|\mathbf{y}) = \int P(\mathbf{y}|\theta, \omega) h(\theta|\omega) d\theta. \quad (22)$$

Again, note that the marginal likelihood does not depend on the unobserved latent variable. Information about the item parameters and the prior distribution of the latent variable (i.e., the structural parameters) is contained in the marginal likelihood. Once more, the optimization of the marginal likelihood provides parameter estimates, standard errors, and model fit statistics.

Numerical Integration

There is, however, one wrinkle. Unlike the multivariate normal factor analysis model, Equation 22 is already in the most simplified form, because the combination of the logistic IRT model and normally distributed latent variable does not lead to a closed-form solution for the integral (the distributions being not conjugate, in Bayesian parlance). Indeed, few practically useful models possess marginal likelihoods that admit the kind of simplifications owing to conjugacy that leads to a closed form expression such as Equation 20. Starting from Bock and Lieberman (1970), likelihood-based item parameter estimation has become a recurring topic in psychometrics because of the numerical challenges of having to integrate over θ .

One popular approach involves numerical quadrature. The integral in Equation 22 may be directly approximated by a sum, where the integrand is evaluated at Q quadrature points X_q , with corresponding weights W_q :

$$\bar{L}(\omega|\mathbf{y}) = \sum_{q=1}^Q \prod_{i=1}^n [T_i(X_q, \omega)]^{y_i} [1 - T_i(X_q, \omega)]^{1-y_i} h(X_q|\omega) W_q. \quad (23)$$

More recent computational algorithms often involve stochastic sampling or alternatives such as variational Bayes (Cai, 2010b; Edwards, 2010; Jeon et al., 2017). We will be discussing these computational approaches in more detail later in this chapter.

Posterior Distribution and Information

Upon estimating the structural parameters from the marginal likelihood, the standard approach in applications of latent variable models for educational assessments is to conduct inferences about those latent variables based on what we have observed. Usually, an empirical Bayes approach is adopted, and the maximum marginal likelihood estimates of the structural parameters $\hat{\omega}$ that we are maximizing the log-posterior with respect to is plugged into the posterior distribution of the latent variables. The conditional independence assumption continues to play an important role in helping derive additional indices from the measurement models that produce numerical summaries of the educational assessment pertaining to its reliability, degree of error and consistency, etc. These topics are important in the practice of educational measurement and are represented in other chapters of this volume in more detail. Here, we rely on the IRT example to illustrate the use of posterior distribution and the derivation of the Fisher information function for a test made up of dichotomously scored items.

From the conditional probability and marginal likelihood in Equations 19 and 22, we can see that the posterior distribution of the latent variable θ is

$$P(\theta | \mathbf{y}, \hat{\omega}) = \frac{\prod_{i=1}^n [T_i(\theta, \hat{\omega})]^{y_i} [1 - T_i(\theta, \hat{\omega})]^{1-y_i} h(\theta | \hat{\omega})}{L(\hat{\omega} | \mathbf{u})} \quad (24)$$

$$\propto \prod_{i=1}^n [T_i(\theta, \hat{\omega})]^{y_i} [1 - T_i(\theta, \hat{\omega})]^{1-y_i} h(\theta | \hat{\omega}),$$

where the second line follows from the fact that the marginal likelihood $L(\hat{\omega} | \mathbf{y})$ at the optimum is a fixed number upon plugging in the structural parameter estimates $\hat{\omega}$. Taking natural logarithms, we see that the log-posterior becomes a sum, due to the factored form of the posterior under conditional independence:

$$\log P(\theta | \mathbf{y}, \hat{\omega}) = \sum_{i=1}^n (y_i \log T_i(\theta, \hat{\omega}) + (1 - y_i) \log (1 - T_i(\theta, \hat{\omega}))) + \log h(\theta | \hat{\omega}). \quad (25)$$

Posterior Summaries

Maximizing the log-posterior in Equation 25 with respect to θ leads to the maximum a posteriori (MAP) scaled score estimator, well known in the practice of test scoring. More precisely, we can write

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \log P(\theta | \mathbf{y}, \hat{\omega}) \quad (26)$$

to indicate that we are maximizing the log-posterior with respect to θ . Omitting the contribution from the population leads to the maximum likelihood estimator, which is

still widely used despite its numerical instability. For example, the maximum likelihood estimator has a tendency to generate extremely low or extremely high estimates of θ (accompanied by undefined standard errors) when the likelihood does not possess a mode. The issue is mitigated with the prior distribution present. The posterior in Equation 24 can also be approximated with quadrature, leading to the following approximating expression for the mean of the posterior distribution:

$$\begin{aligned}\hat{\theta}_{\text{EAP}} &= E(\theta | \mathbf{y}, \hat{\omega}) = \int \theta P(\theta | \mathbf{y}, \hat{\omega}) d\theta \\ &\approx \frac{\sum_{q=1}^Q X_q \prod_{i=1}^n [T_i(X_q, \hat{\omega})]^{y_i} [1 - T_i(X_q, \hat{\omega})]^{1-y_i} h(X_q | \hat{\omega})}{\bar{L}(\omega | \mathbf{u})}.\end{aligned}\quad (27)$$

This is the expected a posteriori (EAP) estimator in test scoring (see Thissen & Wainer, 2001). As an aside, Thissen and Wainer (2001) also extensively discussed the IRT-based characterization of the posterior distribution of θ based on observed summed scores instead of the raw item response patterns. See Cai (2015) and Huang and Cai (2021) for further references, as well as extensions of the approaches discussed in Thissen and Wainer (2001) to the case of multidimensional models.

Information

Fisher's information $F(\theta)$ is a measure of expected (averaged over samples) curvature of the log-likelihood or log-posterior. In other words, it is the expectation of the negative second derivative of Equation 25. Assuming that the population distribution is standard normal, its negative second derivative with respect to θ is 1. After some algebra, this leads to

$$\begin{aligned}F(\theta) &= -E\left(\frac{\partial^2 \log P(\theta | \mathbf{y}, \hat{\omega})}{\partial \theta^2}\right) = \sum_{i=1}^n \beta_i^2 T_i(\theta, \hat{\omega})(1 - T_i(\theta, \hat{\omega})) + 1 \\ &= \sum_{i=1}^n F_i(\theta) + 1,\end{aligned}\quad (28)$$

where $F_i(\theta) = \beta_i^2 T_i(\theta, \hat{\omega})(1 - T_i(\theta, \hat{\omega}))$ is often referred to as the item information function. In other words, we obtained the familiar result that the test's information function is equal to the sum of each item's information function, plus an additional contribution from the population distribution. One may evaluate the Fisher information function and take the square root of its reciprocal (or matrix inverse when θ becomes a vector) as the IRT-based definition of the standard error of measurement (SEM), i.e., $SEM(\theta) = \sqrt{1/F(\theta)}$. We can see that the SEM is a function of the latent variable in IRT. This implies that, to generate the most precise individual estimate, different items should be administered for individuals with different θ values to minimize SEM.

Computerized adaptive testing (e.g., Gibbons et al., 2008; Wainer et al., 2000) research prominently features the use of likelihoods, posteriors, and information functions in the development of optimal item selection algorithms.

The Appraisal of Model Fit

Before one can rely on model-based inference, one must assess the degree of model fit that is relevant to the primary inferential objective at hand. Statements about model fit should always be framed in terms of the sensitivity of model-based inference to the nature of lack of fit. For example, the declaration of adequate model fit for the purpose of dimensionality analysis should not be identical to the needed degree of model fit for the purpose of establishing the stability of anchoring items in a new test form. Neither should one always use the same quantitative measures to gauge model fit. We offer these comments as an antidote to the widespread practice, especially in structural equation modeling, of focusing (and reporting) exclusively on a set of fit measures irrespective of the intended purpose of model fitting.

While we will provide later a more in-depth and technical treatment of the topic of model evaluation, especially in the context of multivariate analysis of discrete/categorical variables frequently encountered in educational measurement, we begin here with a discussion of some general concepts that may be less familiar to educational assessment researchers. Again, it is easiest to use an example. Recall the factor analysis covariance structure model in Equation 20, where y 's covariance matrix is structured as $\Sigma(\omega) = \Lambda\Lambda' + \Psi$, which depends on the free parameters ω specified by a modeler with the parameter matrices (Λ and Ψ). When the number of free parameters in ω is less than the number of nonduplicated elements in Σ , we say that the model is not saturated and has positive degrees of freedom. Note that this patterned covariance structure is a model imposed on the true covariance matrix Σ_0 , which is presumed to exist (and is positive definite). However, there may not be a simple model that can reproduce Σ_0 completely.

Exact and Approximate Goodness of Fit

The first concept is the difference between exact (or absolute) and approximate goodness of fit (Cudeck & Henly, 1991). By goodness of fit, we mean the discrepancy between the model-implied covariance matrix $\Sigma(\omega)$ and the true covariance matrix Σ_0 . This discrepancy can be quantified with a discrepancy function, such as the one motivated by ordinary least squares,

$$F_{OLS}(\Sigma_0, \Sigma(\omega)) = \frac{1}{2} \text{tr}(\Sigma_0 - \Sigma(\omega))^2, \quad (29)$$

which measures the squared difference between the population/true covariance matrix and the model-implied covariance matrix. There are many alternative discrepancy functions (see, e.g., Bollen, 1989), and the one based on normal theory maximum likelihood

(Cudeck & Browne, 1992) is extremely popular in the practice of factor analysis or structural equation modeling:

$$F_{ML}(\Sigma_0, \Sigma(\omega)) = \log|\Sigma(\omega)| + \text{tr}(\Sigma_0 \Sigma(\omega)^{-1}) - \log|\Sigma_0| - n. \quad (30)$$

These discrepancy functions take on a value of 0 if and only if the two inputs Σ_0 and $\Sigma(\omega)$ are identical. Otherwise, they take on positive values that reflect the degree of lack of fit.

The discrepancy function values above are of theoretical interest but cannot be computed directly unless Σ_0 is known, such as in a Monte Carlo simulation. Conceptually, however, their importance cannot be overstated. They indicate that a researcher's model is at best an approximation—a *working model*, as called by MacCallum (2003)—that attempts to capture the unknown true state of affairs. When the working model fails to account for all the nuances of the true state of affairs, even after optimizing over the values of ω , we say that there exists *model error* (MacCallum & Tucker, 1991), which means that there does not exist any parameter value ω in its plausible space of existence that would allow the working model to completely reproduce Σ_0 . The existence of model error is a profound concept and describes what is always true when conducting empirical research, even though the true state of affairs as described by Σ_0 is not known.

The situation is not hopeless, however, because we could still learn a great deal about the unknown structure in Σ_0 by minimizing the discrepancy function. More formally, let

$$\omega_0 = \arg \min_{\omega} F(\Sigma_0, \Sigma(\omega))$$

be the minimizer of the discrepancy, for a given choice of F , the parameters ω_0 that minimize such error is perhaps best described as the *minimum ignorance* estimates along the lines of Akaike (1973) and White (1980, 1982), and M. W. Browne (1984).

The discrepancy function value at the minimizer $F_0 = F(\Sigma_0, \Sigma(\omega_0))$ is 0 if and only if there exists such a set of parameters ω_0 that can completely reproduce the true covariance matrix Σ_0 . In that case, we say the model fits *exactly*. If at the minimizer ω_0 , F_0 is nonzero but small, we say the model fits only *approximately*.

Unfortunately, we do not know the population covariance matrix Σ_0 in general, so F_0 is not directly computable. We do, however, have access to the sample covariance matrix $\mathbf{S}$, which provides (assuming the sampling is unbiased) a consistent estimate of the population covariance matrix. Therefore, we can choose to minimize the discrepancy between the sample covariance matrix and the model, that is,

$$\hat{\omega} = \arg \min_{\omega} F(\mathbf{S}, \Sigma(\omega)).$$

The discrepancy function value at the minimizer $\hat{F} = F(\mathbf{S}, \Sigma(\hat{\omega}))$ signifies both model error and sampling error because the sample size leading to $\mathbf{S}$ is not infinite. This minimized value $\hat{F}$ may be referred to as the *error of approximation*.

Fit Indices

Classical goodness-of-fit tests widely used in moment structure analysis (e.g., Bentler, 1983; M. W. Browne, 1984), and more recently in IRT (e.g., Cai et al., 2006; Maydeu-Olivares & Joe, 2005), are invariably functions of $\tilde{F}$ or of similar measures. Only when model error is absent may we use the standard large sample chi-square approximations to test the absolute fit of the model, with a null hypothesis that says the model fits exactly, that is, no model error. But as Box (1976) observed, “all models are wrong.” We must accept that the “no model error” null hypothesis will always be rejected in the practice of model fitting. It is a fundamentally uninteresting hypothesis because we do not expect (nor do we need) our models to be free of model error to have scientific value. The statistical significance test is therefore routinely a test of whether our sample size is large enough to lead to a rejection of the null hypothesis of exact fit. Hence, other fit measures are needed.

In structural equation modeling, researchers have developed many fit measures that attempt to take the influence of sample size into account or that attempt to quantify the magnitude of model error. Prominent indices include the Tucker–Lewis index (TLI), the root-mean-square error of approximation (RMSEA), and the standardized root-mean-square residual. Guidelines of assessing approximate model fit have been discussed extensively (e.g., Hu & Bentler, 1999; McNeish, & Wolf, 2023; Shi et al., 2019). In IRT, similar measures have been proposed and studied (e.g., Cai et al., 2021; Maydeu-Olivares et al., 2011).

Generalizability Trade-Offs

It is well known to the statistically inclined that a well-fitting model does not guarantee that the model will work as well in an independent replication data set. It has been known for decades (e.g., MacCallum et al., 1992) that seeking good fit alone could lead to models and/or parameters that fit idiosyncratic patterns of a particular data set, due to capitalization on chance. Independently, statisticians have advocated the use of information theory-based indices such as the Akaike information criterion (Akaike, 1973) or Bayesian information criterion (Schwartz, 1978) to penalize models with more parameters that may fit the data better, but potentially at the expense of overfitting. An overfitted model may not generalize (i.e., may not fit as well to new data sampled from the same population as the original). This is an issue that has attracted attention, particularly in the machine learning world (see, e.g., Hastie et al., 2009, Chapter 7). Strategies to mitigate overfitting and improve generalizability range from simple approaches such as cross-validating a model over independent replications (e.g., holding out a random batch of data to test whether the model fits adequately in the hold-out sample) to more formal approaches such as regularization (e.g., the *lasso*; Tibshirani, 1996). It is sufficient to say that even simple approaches such as keeping a random subset of holdout sample for replication, when used routinely in data analysis, can

help significantly improve the generalizability and validity of model-based inference in education.

The Representation of Research Design

An important but often overlooked modeling task in educational assessment is to connect the latent variables representing outcomes with predictor variables, while representing key features of the research design (e.g., randomization) in the model so that unbiased inferences can be drawn. This leads to a blurry boundary between strictly psychometrics-oriented modeling and the models employed in the broader educational statistics field, such as the hierarchical linear model (HLM; Raudenbush & Bryk, 2002). Sometimes the choices of models is clear, such as when group differences at the latent variable and item levels need to be disentangled using differential item functioning methods based on IRT, or when background covariates are entered into a latent regression model, as in NAEP data analysis. Sometimes the choices must be made on a case-by-case basis, such as in the context of using educational assessments to evaluate the efficacy of interventions and programs, when the assessment design itself manifests in increasingly complex dimensionality, or when the data collection design leads to multilevel data structures that make the specification of crossed or nested random effects (latent variables) necessary. In all these cases, it would be important for the modeler to tailor the statistical model to accommodate those aspects of the research design that are the most impactful to the goal of inference. It is helpful to review a prominent example to illustrate our point.

Large-scale educational surveys such as NAEP or Program for International Student Assessment (PISA) provide overall data about educational systems' performance. Multidimensional IRT-based plausible value (PV; Mislevy, 1991; von Davier et al., 2009) methodology is one of the key technical approaches behind these surveys (see also Braun & Kirsch, this volume, for further discussion of PVs in the context of international large-scale assessments). The PV methodology is based on Rubin's multiple imputation theory for missing data analysis (Little & Rubin, 2019; Rubin, 1987). Student achievement latent variables are considered unobserved/missing data. Responses to achievement assessment items and background variables (e.g., demographics, access to resources, motivation, peer context, and school characteristics) are the observed data. To provide efficient estimates of means, variances, and group contrasts at higher levels of aggregation than the individual student, the assessment design makes use of random sampling of items so that each student receives only a small fraction of items selected from the item pool. When individual fractions of missing information are high—and if simple averaging of individual point estimates is used—there is no guarantee that group-level inferences will be unbiased (Mislevy, 1991). The situation is akin to the problem of using single imputation versus multiple imputation in data analysis with missing values present in the data set. In PV methodology, instead of a single point estimate, multiple sets of PVs of student proficiency

variables are generated stochastically from a posterior distribution of the missing data conditional on the entirety of observed data, including background variables (Mislevy, 1991; Mislevy et al., 1992). The imputer's model that generated the PVs combine the background variables with the assessment item response data through a multivariate latent regression model. Therefore, researchers using the PVs to address their own hypotheses do not have to go back to the raw data sets, which are large and complex. Instead, the PVs may be used just as any other multiply imputed data in secondary data analysis.

Suppose NAEP item i assessing domain d has C_i categories, with observed response $c \in \{0, \dots, C_i - 1\}$ from individual j . Let $\mathbf{y}_{jd}$ be the observed student response data in domain d with typical element y_{ijd} denoting item response from individual j to item i in domain d . The individual response vector is $\mathbf{y}_j = (y_{j1}, \dots, y_{jD})$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ collects together all observed outcomes. Of interest to the assessment developers and users is estimating the unobserved student achievement in the D domains, and the vector containing these latent achievement variables for student j is represented by $\boldsymbol{\theta}'_j = (\theta_{j1}, \dots, \theta_{jd}, \dots, \theta_{jD})$. In NAEP, the domains are correlated. Finally, let $\mathbf{X} = (x_1, \dots, x_j, \dots, x_N)$ be the matrix of background variables, where $\mathbf{x}_j$ is the covariate vector for individual j , and let N denote the total sample size.

Next, in the conditioned latent regression model that produces the PVs, the latent proficiency vector $\boldsymbol{\theta}_j$, is regressed on $\mathbf{x}_j$

$$\boldsymbol{\theta}_j = \boldsymbol{\Gamma} \mathbf{x}_j + \boldsymbol{\varepsilon}_j,$$

where the disturbance term $\boldsymbol{\varepsilon}_j$ follows a D -variate normal distribution with zero means and an unstructured covariance matrix $\boldsymbol{\Sigma}$. The multivariate regression model implies that $\boldsymbol{\theta}_j$ is D -variate normally distributed with conditional mean vector $\boldsymbol{\Gamma} \mathbf{x}_j$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Gamma}$ represents a matrix of unknown regression coefficients (see Thomas & Gan, 1997). The background covariates contribute to the explanation of the individual-specific outcomes by giving each student potentially different conditional means. Let $f(\boldsymbol{\theta}_j; \boldsymbol{\Gamma} \mathbf{x}_j, \boldsymbol{\Sigma})$ represent a multivariate normal density function with mean $\boldsymbol{\Gamma} \mathbf{x}_j$ and covariance matrix $\boldsymbol{\Sigma}$.

We now need the conditional independence assumption. Suppose that the items in each of the proficiency domains are independent once the latent proficiency θ_{jd} is accounted for and that each item belongs to only one domain, that is, the independent cluster pattern in factor analysis. The posterior distribution of the latent proficiencies $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$ is then proportional to the following joint likelihood:

$$f(\boldsymbol{\Theta} | \mathbf{X}, \mathbf{Y}) \propto \prod_{j=1}^N f(\boldsymbol{\theta}_j; \boldsymbol{\Gamma} \mathbf{x}_j, \boldsymbol{\Sigma}) \prod_{d=1}^D f_d(\mathbf{y}_{jd} | \theta_{jd}), \quad (31)$$

from which sets of PVs of $\boldsymbol{\Theta}$ can be drawn. The main lesson here is that the model described in Equation 31 is custom built to encode core features of the assessment design, which should be a desirable goal in modeling of educational assessment data.

A GENERAL FRAMEWORK

Given the foregoing discussions, we present a general multivariate statistical framework that combines elements of factor analysis, multidimensional, multilevel (including cross-classified relations), diagnostic classification IRT models, with conditioning covariates and predictors. The model may be constrained to obtain existing, known factor analysis or IRT models, as Table 10.1 illustrates. From this framework, one can obtain more specialized models that are fit for particular research designs or inferential purposes akin to the practice of model building in the field of structural equation modeling.

Measurement Models

The Nested Case

To begin, let us consider the case of fully nested relationship in our data collection design, using the language of HLM. Suppose the observed data are from a Level 1 unit j nested within independent sampling unit (Level 2 unit) k , for example, a student nested within a school. For a generic item/observed variable i , we may denote the linear combination of latent variables as

$$\eta_{ijk} = \beta'_{i1} \begin{pmatrix} h_{i1}(\mathbf{Q}_1, \boldsymbol{\xi}_{jk}) \\ \boldsymbol{\theta}_{jk} \end{pmatrix} + \beta'_{i2} \begin{pmatrix} h_{i2}(\mathbf{Q}_2, \boldsymbol{\xi}_k) \\ \boldsymbol{\theta}_k \end{pmatrix}, \quad (32)$$

where β_{i1} and β_{i2} are the item slope parameters, with $\mathbf{Q}_1$ and $\mathbf{Q}_2$ the $\mathbf{Q}$ matrices at Levels 1 and 2, respectively. The subscripts on $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ indicate whether the latent variables vary over Level 1 units (j and k) or over Level 2 units (k only). The dimensions of the slope vectors are dependent on both the number of continuous latent dimensions (factors) in $\boldsymbol{\theta}$ and the number of terms, such as interaction effects, generated by the h functions and $\mathbf{Q}$ matrices at each level for the diagnostic attributes in $\boldsymbol{\xi}$. We will assume that there are K Level 2 units in total, and each Level 2 unit consists of N_k Level 1 units. The total number of Level 1 units is $N = \sum_k N_k$.

The Cross-Classified Case

In educational research, we also frequently encounter cases in which the lowest unit of data collection design belongs to two or more nonnested higher level units simultaneously. This yields a cross-classified data structure (e.g., Beretvas, 2011; W. J. Browne et al., 2001; Fielding & Goldstein, 2006; Goldstein, 1994; Huang, 2021; Jeon et al., 2017; Raudenbush & Bryk, 2002). When assessment instruments made up of multiple items are employed to measure unobserved variables, we have multivariate categorical outcome variables with a cross-classified data structure. An example is the ubiquitous use of student surveys for evaluation of instruction in higher education. In educational assessment, the person-by-item response matrix may be perceived of as a form of cross-classification as well. This leads directly to the explanatory and random item models (see de Boeck & Wilson, 2004) that promise to significantly expand the scope of the linear variance components approach (Searle et al., 2006)

that remains to this day foundational in generalizability theory (G theory; Shavelson & Webb, 1991). These random-item models also influenced new developments in models for network psychometrics (e.g., Jin & Jeon, 2019).

With model building in mind, let us consider adaptations to the linear combination in Equation 32. For item i that is within a cell jk defined by a two-way cross-classification, the linear combination of latent variables can take the following form:

$$\eta_{ijk} = \beta'_{i1} \begin{pmatrix} h_{i1}(\mathbf{Q}_1, \boldsymbol{\xi}_j) \\ \boldsymbol{\theta}_j \end{pmatrix} + \beta'_{i2} \begin{pmatrix} h_{i2}(\mathbf{Q}_2, \boldsymbol{\xi}_k) \\ \boldsymbol{\theta}_k \end{pmatrix}, \quad (33)$$

where the $\boldsymbol{\xi}_j$ and $\boldsymbol{\theta}_j$ vary over the first facet of the cross-classification, and $\boldsymbol{\xi}_k$ and $\boldsymbol{\theta}_k$ vary over the second facet. They may be aptly termed the row side and the column side in an imaginary two-way table wherein each cell contains multivariate outcomes that are influenced by the latent variables. For example, the rows may represent students and the columns the courses they took—or perhaps the rows are essays and the columns are the raters. We assume that there are J rows and K columns, with the total number of cells equal to $N = J \times K$.

Link Functions and Conditional Distributions

Substituting the linear combination in Equations 32 or 33 into, for example, the 2PL model, results in the item's conditional response probability, for example,

$$T_i(1 | \eta_{ijk}) = \frac{1}{1 + \exp[-(\alpha_i + \eta_{ijk})]}, \quad (34)$$

with an additional item intercept parameter α_i . For an item with multiple categorical responses $c \in \{0, \dots, C_i - 1\}$, the conditional probability of observing response y_{ijk} is a multinomial with trial size 1 in C_i categories:

$$f_{\omega}(y_{ijk} | \eta_{ijk}) = \prod_{c=0}^{C_i-1} T_i(c | \eta_{ijk})^{\chi_c(y_{ijk})}, \quad (35)$$

where the indicator function $\chi_c(y_{ijk})$ takes a value of 1 if and only if $y_{ijk} = c$ and is zero otherwise. We use the subscript ω to emphasize that the conditional distribution depends on a set of structural parameters contained in this vector.

The conditional probability $T_i(c | \eta_{ijk})$ may take on a variety of forms. The 2PL, 3PL, and GPC models presented earlier are popular candidates. Samejima's (1969) graded response model with cumulative logits is another practical choice. Its item conditional probability is usually written as

$$T_i(c | \eta_{ijk}) = \frac{1}{1 + \exp[-(\alpha_{i,c} + \eta_{ijk})]} - \frac{1}{1 + \exp[-(\alpha_{i,c+1} + \eta_{ijk})]}, \quad (36)$$

to indicate that the category probability is the difference between adjacent cumulative probabilities, which are themselves modeled after the 2PL model. The two boundary cases are defined as

$$T_i(0|\eta_{ijk}) = 1 - \frac{1}{1 + \exp[-(\alpha_{i,1} + \eta_{ijk})]},$$

and

$$T_i(C_i - 1|\eta_{ijk}) = \frac{1}{1 + \exp[-(\alpha_{i,C_i-1} + \eta_{ijk})]} - 0,$$

for the obvious reasons that the cumulative probability of the response falling in category 0 or above is equal to 1.0 and that there is simply no probability for a response coded as C_i beyond the highest response of $C_i - 1$. The graded model subsumes the 2PL as a special case (i.e., when $C_i = 2$).

As an aside, we can handle the linear factor analysis (continuous outcome) case easily by assuming that the conditional distribution of the outcome variable is normal,

$$f_{\omega}(y_{ijk}|\eta_{ijk}) \sim N(\alpha_i + \eta_{ijk}, \sigma_i^2),$$

where α_i is the usual measurement intercept parameter. If the outcome is a count variable, we can implement a log-linear model with a Poisson conditional distribution,

$$f_{\omega}(y_{ijk}|\eta_{ijk}) \sim \text{Pois}(\alpha_i + \eta_{ijk}).$$

Other link functions and conditional distributions may be assumed if warranted, such as log-normal distributions for response time data (van der Linden, 2007). One simply has to follow standard theory in generalized linear mixed modeling (e.g., McCulloch et al., 2008) and in generalized latent variable modeling (e.g., Bartholomew et al., 2011; Cai, 2008; Skrondal & Rabe-Hesketh, 2004).

Conditioning on Predictors

To build the model further, we invoke the conditional independence assumption, so that the conditional response pattern probability for the response vector $\mathbf{y}_{jk} = (y_{1jk}, \dots, y_{njk})$ containing n observations is expressed as a product of individual item response probabilities. For the nested case, we have

$$f_{\omega}(\mathbf{y}_{jk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \prod_{i=1}^n f_{\omega}(y_{ijk}|\boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k). \quad (37)$$

For the cross-classified case, we have

$$f_{\omega}(\mathbf{y}_{jk}|\boldsymbol{\xi}_j, \boldsymbol{\xi}_k, \boldsymbol{\eta}_j, \boldsymbol{\eta}_k) = \prod_{i=1}^n f_{\omega}(y_{ijk}|\boldsymbol{\xi}_j, \boldsymbol{\xi}_k, \boldsymbol{\eta}_j, \boldsymbol{\eta}_k). \quad (38)$$

The presence of both types of latent variables (continuous and discrete) at both levels allows the model to be more flexible than standard multidimensional or multilevel IRT models, as well as DCMs. The remaining task involves the conditioning of the $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ variables by regressing them on covariates or predictors.

In general, the conditioning regression model is a multivariate linear model that takes the following form:

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\varepsilon},$$

where the equation disturbance term $\boldsymbol{\varepsilon}$ is typically assumed to be multivariate normal with zero means and an error covariance matrix that may contain estimable parameters.

Consider the latent regression for the nested case, where $\mathbf{x}_{jk}$ indicates a vector of background covariates or predictors that varies over Level 1 units and, if applicable, $\mathbf{x}_k$ is a vector of Level 2 predictors. There are two simultaneous regression equations:

$$\begin{pmatrix} \boldsymbol{\eta}_{jk} \\ \boldsymbol{\eta}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{jk} \\ \mathbf{x}_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{jk} \\ \boldsymbol{\varepsilon}_k \end{pmatrix}, \quad (39)$$

where the lower left block of the regression coefficient matrix is zero because Level 2 latent variables should not be regressed on Level 1 predictors. A set of similar, and simpler, equations can be specified for the cross-classified case:

$$\begin{pmatrix} \boldsymbol{\eta}_j \\ \boldsymbol{\eta}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{x}_j \\ \mathbf{x}_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_j \\ \boldsymbol{\varepsilon}_k \end{pmatrix}, \quad (40)$$

where the predictors vary along the same facets of the latent variables, that is, row-side latent variables are regressed on row-side predictors and the same for the column side. In both nested and cross-classified cases, the predictor vectors may contain a constant term so that the regression intercept parameters may be estimated.

Likelihoods

Combining various parts reviewed earlier, and following standard modeling practices, the latent variables can now be integrated out of the joint likelihood. Suppose $\boldsymbol{\omega}$ is the $d \times 1$ vector that collects together all free parameters in the model. These include parameters from all the items (the α 's, β 's, and γ 's), parameters for the latent variables, the regression coefficients, and covariance matrices of the error terms.

The Nested Case

For the nested case, the integration is done sequentially. In the first step, the Level 1 discrete latent attributes are integrated out:

$$f_{\omega}(\mathbf{y}_{jk} | \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) = \int f_{\omega}(\mathbf{y}_{jk} | \boldsymbol{\xi}_{jk}, \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) f_{\omega}(\boldsymbol{\xi}_{jk}) d\boldsymbol{\xi}_{jk}, \quad (41)$$

where $f_{\omega}(\boldsymbol{\xi}_{jk})$ is the prior distribution on the Level 1 attributes. Second, the Level 1 continuous latent variables are integrated out, bringing along the conditioning covariates as well:

$$f_{\omega}(\mathbf{y}_{jk} | \boldsymbol{\xi}_k, \boldsymbol{\eta}_k, \mathbf{x}_{jk}, \mathbf{x}_k) = \int f_{\omega}(\mathbf{y}_{jk} | \boldsymbol{\xi}_k, \boldsymbol{\eta}_{jk}, \boldsymbol{\eta}_k) f_{\omega}(\boldsymbol{\eta}_{jk} | \mathbf{x}_{jk}, \mathbf{x}_k) d\boldsymbol{\eta}_{jk}, \quad (42)$$

where $f_{\omega}(\eta_{jk} | \mathbf{x}_{jk}, \mathbf{x}_k)$ follows from the first term of the matrix equations in Equation 39. In the next step, one requires the assumption of conditional independence again, this time of the independence of N_k Level 1 response vectors nested within Level 2 unit k , conditionally on all the Level 2 latent variables:

$$\begin{aligned} f_{\omega}(\{\mathbf{y}_{jk}\}_{j=1}^{N_k} | \{\mathbf{x}_{jk}\}_{j=1}^{N_k}, \mathbf{x}_k) \\ = \int \int \prod_{j=1}^{N_k} f_{\omega}(\mathbf{y}_{jk} | \xi_k, \eta_k, \mathbf{x}_{jk}, \mathbf{x}_k) f_{\omega}(\xi_k) f_{\omega}(\eta_k | \mathbf{x}_k) d\xi_k d\eta_k, \end{aligned} \quad (43)$$

where $f_{\omega}(\xi_k)$ is the prior distribution on the Level 2 attributes and $f_{\omega}(\eta_k | \mathbf{x}_k)$ follows from the second row of the matrix equation in Equation 39. The parenthetical notation $\{\mathbf{y}_{jk}\}_{j=1}^{N_k}$ and $\{\mathbf{x}_{jk}\}_{j=1}^{N_k}$ collects all the Level 1 outcome and covariate observations within Level 2 unit k . Suffice it to say that Equation 43 contains the marginal likelihood function of all parameters of the model. In both the nested and the cross-classified cases, we shall collect all the outcome variables into $\mathbf{Y}$ and all predictors into $\mathbf{X}$.

Summing across the (assumed) independent Level 2 units, the overall marginal log-likelihood function is

$$\log L(\omega | \mathbf{Y}, \mathbf{X}) = \sum_{k=1}^K \log f_{\omega}(\{\mathbf{y}_{jk}\}_{j=1}^{N_k} | \{\mathbf{x}_{jk}\}_{j=1}^{N_k}, \mathbf{x}_k). \quad (44)$$

The Cross-Classified Case

For the cross-classified case, the integration is iterated over all rows and columns in the two-way cross-classification to arrive at the marginal likelihood function:

$$\begin{aligned} L(\omega | \mathbf{Y}, \mathbf{X}) = \prod_{k=1}^K \int \int \prod_{j=1}^J \int \int f_{\omega}(\mathbf{y}_{jk} | \xi_j, \eta_j, \xi_k, \eta_k) f_{\omega}(\xi_j) \\ f_{\omega}(\eta_j | \mathbf{x}_j) d\xi_j d\eta_j f_{\omega}(\xi_k) f_{\omega}(\eta_k | \mathbf{x}_k) d\xi_k d\eta_k. \end{aligned} \quad (45)$$

We begin by conditioning on a particular row j 's latent variables. We realize that if we further condition on the column-side latent variables, all columns become conditionally independent within this row. Next, with the regression equation in Equation 40, we can integrate out the continuous latent variables and bring the predictors into the conditioning set. The marginal log-likelihood is the logarithm of Equation 45. Software such as flexMIRT (Houts & Cai, 2022) implements nearly all of the features of the generalized modeling framework presented previously and estimates the structural parameters with maximum marginal likelihood. Alternatives such as fully Bayesian approaches may also be adopted.

ESTIMATION METHODS

At this point, with the marginal likelihoods fully specified, the task in front of the statistical modeler is that of the estimation of the structural parameters of the model in ω and the appraisal of model fit, before inferences about the latent variables are drawn using the empirical Bayes approach already discussed (see, e.g., Equation 20), which

is about test scoring in the practice of educational measurement with comprehensive discussions appearing elsewhere (e.g., Thissen & Wainer, 2001).

The optimization of the marginal log-likelihoods in Equations 44 or 47 typically requires advanced iterative numerical methods because closed-form solutions are unavailable. Various computational approaches may be used, including the venerable expectation–maximization (EM) algorithm with fixed quadrature (see Equation 19) due to Bock and Aitkin (1981), or the more recent Metropolis–Hastings Robbins–Monro algorithm (MH-RM; Cai, 2008, 2010b, 2010c) based on stochastic approximation. Alternatively, one may also adopt fully Bayesian estimation using Markov chain Monte Carlo (MCMC) sampling methods (Hastings, 1970; Metropolis et al., 1953) for maximal flexibility in model specification and estimation. See Edwards (2010) for a review in the context of IRT or borrow an optimization toolkit from machine learning (Urban & Bauer, 2021). For a recent review, please see Y. Chen et al. (2022).

Despite the algorithmic advances now at the fingertips of modelers of educational assessment data, it is important to review some computational concepts so that modelers may be equipped with the knowledge to diagnose computational troubles that may frustrate any data analyses involving iterative optimization of models and parameters. We begin with the missing data formulation present in Dempster et al.’s (1977) classic paper that coined the term expectation–maximization algorithm.

For now, let us return to the basic case of a unidimensional IRT model as in Equation 13 with a standard normal prior/population distribution. Let $\mathbf{Y}$ be a matrix of all observed item responses, and $\mathbf{y}_j$ is the $n \times 1$ vector of responses from individual j . The marginal likelihood function for all the structural parameters in ω can be expressed as

$$\begin{aligned} L(\omega | \mathbf{Y}) &= \prod_{j=1}^N \int \prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}} h(\theta) d\theta \\ &= \prod_{j=1}^N \log L(\omega | \mathbf{y}_j). \end{aligned} \quad (46)$$

Because the marginal likelihood does not depend on the unobserved θ values, it may also be referred to as the *observed* data likelihood. Implicit in the “observed data” reference is a realization that the latent variables are the missing values. Suppose θ is a vector that collects together all N latent variable scores and $\mathbf{y}_i$ is the $N \times 1$ vector of responses to item i . If we were able to observe the latent variables, then the “complete data” likelihood function of item parameters takes on a factored form, again by virtue of the conditional independence assumption of item responses and the independence of individuals. We can express it as

$$\begin{aligned} L(\omega | \mathbf{Y}, \theta) &= \left[\prod_{j=1}^N h(\theta_j) \right] \left[\prod_{j=1}^N \prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}} \right] \\ &\propto \prod_{j=1}^N L(\omega | \mathbf{y}_j, \theta_j) = \prod_{i=1}^n L(\omega | \mathbf{y}_i, \theta), \end{aligned} \quad (47)$$

where

$$L(\omega | \mathbf{y}_j, \theta_j) = \prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}}$$

is the complete data likelihood for individual j , and

$$L(\omega | \mathbf{y}_i, \theta) = \prod_{j=1}^N \prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}}$$

is the complete data likelihood for item i . The proportionality on the second line of Equation 47 holds because $h(\theta_j)$ is assumed to be standard normal and becomes a constant if the θ 's were known. Equation 47 makes it clear that the complete data likelihood is a constant multiple of the product of n item-specific complete data likelihoods or N individual-specific complete data likelihoods.

Of course, we do not observe the latent variable θ values. But the posterior distribution of the latent variables given the observed item responses provides the basis of a principled way to fill in (impute) the missing data. For an individual, recall the posterior is defined as the following (Equation 20):

$$P(\theta | \mathbf{y}_j, \omega) = \frac{\prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}} h(\theta | \omega)}{\int \prod_{i=1}^n [T_i(\theta, \omega)]^{y_{ij}} [1 - T_i(\theta, \omega)]^{1-y_{ij}} h(\theta | \omega) d\theta}. \quad (48)$$

The distribution function itself is analytically intractable, but the posterior is proportional to the joint distribution in the numerator, which implies that with MCMC, the posterior can be sampled, for example, with Patz and Junker's (1999) algorithm. But even more important, we can verify that the following equality holds under mild regularity conditions:

$$\frac{\partial \log L(\omega | \mathbf{y}_j)}{\partial \omega} = \int \frac{\partial \log L(\omega | \mathbf{y}_j, \theta_j)}{\partial \omega} P(\theta_j | \mathbf{y}_j, \omega) d\theta_j. \quad (49)$$

Equation 49 is known as Fisher's identity (Fisher, 1925). This is arguably the second most important computational concept in latent variable modeling, after the missing data formulation. Fisher's identity states that the gradient of the observed data log-likelihood $\log L(\omega | \mathbf{y}_j)$ is equal to the conditional expectation of the gradient of the complete data log-likelihood $\log L(\omega | \mathbf{y}_j, \theta_j)$. The expectation is taken over the posterior distribution of the latent variables given the observed variables. This powerful insight suggests that instead of trying to maximize the observed data likelihood, which is direct but often more difficult, one should consider iteratively maximizing the conditional expected complete data likelihood (i.e., the right-hand side of Equation 19), which is indirect but more tractable. It lies behind Bock and Aitkin's (1981) EM algorithm and motivated many modern variants, including the MH-RM algorithm.

Bock–Aitkin EM Algorithm

From Equation 19, we see that the marginal likelihood can be approximated to arbitrary precision by replacing the integration with a summation over a set of Q quadrature points X_q spread over θ , with quadrature weights W_q . Bock and Aitkin (1981) saw that the height of the posterior at quadrature point X_q can be approximated to arbitrary precision as well:

$$P(X_q | \mathbf{y}_j, \omega) \approx \frac{[T_i(X_q, \omega)]^{y_{ij}} [1 - T_i(X_q, \omega)]^{1-y_{ij}} W_q}{\sum_{q=1}^Q \prod_{i=1}^n [T_i(X_q, \omega)]^{y_{ij}} [1 - T_i(X_q, \omega)]^{1-y_{ij}} h(X_q) W_q}. \quad (50)$$

Ignoring constants and taking the logarithm of Equation 47, the complete data log-likelihood is

$$\begin{aligned} \log L(\omega | \mathbf{Y}, \theta) &= \sum_{j=1}^N \sum_{i=1}^n y_{ij} \log T_i(\theta_j, \omega) \\ &\quad + \sum_{j=1}^N \sum_{i=1}^n (1 - y_{ij}) \log(1 - T_i(\theta_j, \omega)). \end{aligned} \quad (51)$$

Following Fisher's identity, the conditional expected complete data likelihood given provisional item parameter values ω^* can also be approximated by quadrature,

$$\begin{aligned} M(\omega | \mathbf{Y}, \omega^*) &\approx \sum_{j=1}^N \sum_{q=1}^Q \sum_{i=1}^n y_{ij} \log T_i(X_q, \omega) P(X_q | \mathbf{y}_j, \omega^*) \\ &\quad + \sum_{j=1}^N \sum_{q=1}^Q \sum_{i=1}^n (1 - y_{ij}) \log(1 - T_i(X_q, \omega)) P(X_q | \mathbf{y}_j, \omega^*) \end{aligned} \quad (52)$$

By interchanging the order of summation, Bock and Aitkin (1981) realized that the posterior probabilities can be accumulated over individuals first:

$$\begin{aligned} M(\omega | \mathbf{Y}, \omega^*) &\approx \sum_{i=1}^n \sum_{q=1}^Q \log T_i(X_q, \omega) \left(\sum_{j=1}^N y_{ij} P(X_q | \mathbf{y}_j; \omega^*) \right) \\ &\quad + \sum_{i=1}^n \sum_{q=1}^Q \log(1 - T_i(X_q, \omega)) \left(\sum_{j=1}^N (1 - y_{ij}) P(X_q | \mathbf{y}_j; \omega^*) \right) \\ &= \sum_{i=1}^n \sum_{q=1}^Q r_{iq} \log T_i(X_q, \omega) + \sum_{i=1}^n \sum_{q=1}^Q \bar{r}_{iq} \log(1 - T_i(X_q, \omega)), \end{aligned} \quad (53)$$

where $r_{iq} = \sum_{j=1}^N y_{ij} P(X_q | \mathbf{y}_j; \omega^*)$ can be understood as the expected number of positive/correct/endorsement responses to item i at quadrature point X_q , and

$\bar{r}_{iq} = \sum_{j=1}^J (1 - y_{ij}) P(X_q | \mathbf{y}_j; \boldsymbol{\omega}^*)$ is its counterpart for the negative/incorrect/disagreement responses. Taken together, let $n_{iq} = r_{iq} + \bar{r}_{iq} = \sum_{j=1}^J P(X_q | \mathbf{y}_j; \boldsymbol{\omega}^*)$ be the expected number of individuals at quadrature point X_q . We then have

$$M(\boldsymbol{\omega} | \mathbf{Y}, \boldsymbol{\omega}^*) \approx \sum_{i=1}^n \sum_{q=1}^Q r_{iq} \log T_i(X_q, \boldsymbol{\omega}) + (n_{iq} - r_{iq}) \log(1 - T_i(X_q, \boldsymbol{\omega})). \quad (54)$$

Equation 54 implies that the conditional expected complete data log-likelihood is a set of n parallel logistic regression log-likelihoods for “grouped” data, with the quadrature points X_q serving as the predictor values and weights given by n_{iq} and r_{iq} . Optimizing $M(\boldsymbol{\omega} | \mathbf{Y}, \boldsymbol{\omega}^*)$ leads to new and improved item parameter estimates, which can be used to update the posterior probabilities, and consequently the weights n_{iq} and r_{iq} .

In general, Bock–Aitkin EM (or any EM algorithm) alternates between the following two steps from a set of initial parameter estimates, say $\boldsymbol{\omega}^{(0)}$, and it generates a sequence of parameter estimates $\boldsymbol{\omega}^{(0)}, \dots, \boldsymbol{\omega}^{(k)}$ that converges under some very general conditions to the maximum marginal likelihood estimate of $\boldsymbol{\omega}$ as the number of cycles k tends to infinity (Wu, 1983).

E-Step

Given $\boldsymbol{\omega}^{(k)}$, evaluate the conditional expected complete data log-likelihood $M(\boldsymbol{\omega} | \mathbf{Y}; \boldsymbol{\omega}^{(k)})$, which is taken as a function of $\boldsymbol{\omega}$.

M-Step

Maximize $M(\boldsymbol{\omega} | \mathbf{Y}; \boldsymbol{\omega}^{(k)})$ to yield updated parameter estimates $\boldsymbol{\omega}^{(k+1)}$. Go back to E-step and repeat until estimates from adjacent cycles stabilize (i.e., convergence).

MH-RM Algorithm

Although Bock–Aitkin EM is highly efficient for unidimensional IRT parameter estimation, it does not generalize well to the case of multidimensional IRT because of the inherent limitation of numerical quadrature. The likelihoods and posteriors are evaluated over grids of quadrature points, leading to exponentially increasing computational burden as the number of latent variables increases. This is called the *curse of dimensionality* in the literature. As educational assessments become more complex, alternative estimation approaches are needed.

Cai (2008, 2010a, 2010b) reasoned that if one can randomly draw plausible values or imputations of θ from the posterior $P(\theta | \mathbf{y}_j, \boldsymbol{\omega}^*)$, with provisional parameter estimates $\boldsymbol{\omega}^*$, the left-hand side of Equation 49 can be approximated by Monte Carlo integration, that is,

$$\frac{\partial \log L(\boldsymbol{\omega} | \mathbf{y}_j)}{\partial \boldsymbol{\omega}} \approx \frac{1}{V} \sum_{v=1}^V \frac{\partial \log L(\boldsymbol{\omega} | \mathbf{y}_j, \theta_{jv})}{\partial \boldsymbol{\omega}}, \quad (55)$$

where θ_{jv} are random draws from $P(\theta \mid \mathbf{y}_j, \boldsymbol{\omega}^*)$. Stated differently, the Monte Carlo average of the gradients of the complete data log-likelihood contain (approximately) the same ascent direction as the observed data log-likelihood gradients. These random samples can be easily produced with the help of an MCMC sampling method, such as the Metropolis–Hastings algorithm. At this point, Cai (2008) drew an analogy to the engineering applications of Robbins and Monro's (1951) classical stochastic approximation method. The approximate ascent directions are corrupted by random sampling noise, but it can be gradually filtered out in the Robbins–Monro (RM) recursions with the use of a sequence of decreasing gain constants to achieve a point-wise convergent algorithm (see Cai, 2010a, for a proof) that resembles the EM algorithm. Thus, the number of random samples V can be a fixed and small number (as low as 1) throughout all the iterations of the algorithm.

In general, starting from $k = 0$ (initial values), cycle $k + 1$ of the MH-RM algorithm consists of three steps:

Imputation

Given provisional parameter estimates $\boldsymbol{\omega}^{(k)}$, V sets of random samples of the latent variables $\boldsymbol{\theta}_v^{(k+1)}$ are drawn. For each individual, the draws may come from a Metropolis–Hastings sampler that has, as its unique invariant (target) distribution, the posterior distribution $P(\theta_j \mid \mathbf{y}_j, \boldsymbol{\omega}^{(k)})$ of the missing data given the observed data and provisional parameter values. Form complete data sets as $\{\mathbf{Y}, \boldsymbol{\theta}_v^{(k+1)}\}_{v=1}^V$.

Approximation

In the second step, based on the imputed data, the complete data log-likelihood and its derivatives are evaluated so that the ascent directions for parameters can be determined. The complete data gradient (score function) is approximated as

$$\mathbf{s}_{k+1} = \frac{1}{V} \sum_{v=1}^V \frac{\partial \log L(\boldsymbol{\omega}^{(k)} \mid \mathbf{Y}, \boldsymbol{\theta}_v^{(k+1)})}{\partial \boldsymbol{\omega}}. \quad (56)$$

At the same time, to improve stability and speed, we also evaluate a Monte Carlo approximation to the conditional expected complete data information matrix:

$$\mathbf{I}_{k+1} = -\frac{1}{V} \sum_{v=1}^V \frac{\partial^2 \log L(\boldsymbol{\omega}^{(k)} \mid \mathbf{Y}, \boldsymbol{\theta}_v^{(k+1)})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}. \quad (57)$$

Because of conditional independence, the complete data information matrix is often block diagonal, vastly reducing the number of elements that must be computed.

Robbins–Monro Update

In the third step, Robbins–Monro stochastic approximation filters are applied when updating the estimates of item parameters. First, the Robbins–Monro filter is applied

to obtain a recursive stochastic approximation of the conditional expectation of the complete data information matrix

$$\mathbf{I}_{k+1} = \mathbf{I}_k + g_k (\mathbf{I}_{k+1} - \mathbf{I}_k), \quad (58)$$

where g_k is a sequence of nonnegative *gain constants* such that $g_k \in (0, 1]$, $\sum_{k=0}^{\infty} g_k = \infty$, $\sum_{k=0}^{\infty} g_k^2 < \infty$. Next, we use the Robbins–Monro filter again when updating the parameters:

$$\boldsymbol{\omega}^{(k+1)} = \boldsymbol{\omega}^{(k)} + g_k (\mathbf{I}_{k+1})^{-1} \mathbf{s}_{k+1}. \quad (59)$$

The iterations are started from some initial parameter values $\boldsymbol{\omega}^{(0)}$ and terminated when the estimates stabilize. Cai (2008, 2010a) showed that the sequence of parameters converges with probability 1 to a local maximum of the marginal/observed data likelihood function $L(\boldsymbol{\omega} | \mathbf{Y})$. Typically, the sequence of gain constants are taken to be $g_k = 1/(k+1)$, in which case the initial choice of $\mathbf{I}_0$ becomes arbitrary. Cai (2010a) provided formulas for recursively approximating the parameter error covariance matrix, as well as further discussions about convergence checking. The Metropolis-Hastings samplers can be tuned for optimal efficiency using the theory and approach developed by Robert & Rosenthal (2001).

THEORY AND METHODS FOR MODEL FIT EVALUATION

After obtaining estimates of the structural parameters and uncertainty estimates such as standard errors, the immediate task is to evaluate the fit of the model. In the context of factor analysis and structural equation modeling, this topic has been thoroughly discussed and is part of the standard practice. In the application of statistical modeling of educational assessment data, however, discussions of a coherent theory and methods of model fit evaluation have been scarce because discrete multivariate analysis is not a widely introduced topic. In this section, we will review some important concepts particularly relevant to the applications of models for discrete outcome variables (e.g., IRT and DCM).

The Multinomial Formulation

Let there be n items in a test. Suppose we use the model in Equation 6. The IRT model specifies the marginal probability of the item response pattern as the following integral:

$$P(\mathbf{y} | \boldsymbol{\omega}) = \int P(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\omega}) h(\boldsymbol{\theta} | \boldsymbol{\omega}) d\boldsymbol{\theta} = \pi_{\mathbf{y}}(\boldsymbol{\omega}). \quad (60)$$

The parenthetical notation $\pi_{\mathbf{y}}(\boldsymbol{\omega})$ in Equation 60 is used to emphasize the fact that it is the model. The marginal probability depends on the response pattern $\mathbf{y}$, the $d \times 1$ vector of structural parameters $\boldsymbol{\omega}$, the conditional probability $P(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\omega})$, and the latent variable distribution $h(\boldsymbol{\theta} | \boldsymbol{\omega})$.

Recall that C_i is the number of categories for item i . The total number of possible item response patterns is $C = \prod_{i=1}^n C_i$. This can be arranged (grouped) in the form of a contingency table under multinomial sampling. Based on a sample of N respondents, let the observed proportion associated with pattern $\mathbf{y}$ be denoted p_y . The sampling model for this contingency table is a multinomial distribution with C cells and N trials. The log-likelihood for the item parameters ω is proportional to

$$\log L(\omega) \propto N \sum_y p_y \log \pi_y(\omega), \quad (61)$$

where the summation runs over all C possible response patterns if all p_y are positive. Otherwise, the summation can be assumed to run over the observed response patterns when some observed p_y values are zero. Maximization of the log-likelihood leads to the maximum marginal likelihood estimator $\hat{\omega}$.

Upon finding $\hat{\omega}$, the model generates model-implied probabilities for each response pattern $\pi_y(\hat{\omega}) = \hat{\pi}_y$. Suppose the model-implied response pattern probabilities $\hat{\pi}_y$ are collected into a $C \times 1$ vector $\hat{\pi}$ of all model-implied response pattern probabilities. By analogy, let a $C \times 1$ vector π contain the true response pattern probabilities. The observed proportions p_y can be collected into a $C \times 1$ vector $\mathbf{p}$ as well. For example, for three dichotomously items, there are $2^3 = 8$ item response patterns, and the response pattern probabilities and observed proportions are

$$\pi = \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \pi_{111} \end{pmatrix}, \quad \hat{\pi} = \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix} = \begin{pmatrix} \pi_{000}(\hat{\omega}) \\ \pi_{001}(\hat{\omega}) \\ \pi_{010}(\hat{\omega}) \\ \pi_{011}(\hat{\omega}) \\ \pi_{100}(\hat{\omega}) \\ \pi_{101}(\hat{\omega}) \\ \pi_{110}(\hat{\omega}) \\ \pi_{111}(\hat{\omega}) \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{pmatrix}. \quad (62)$$

Distribution of Residuals

Based on standard results from discrete multivariate analysis (Bishop et al., 1975), the maximum likelihood estimator $\hat{\omega}$ is consistent, asymptotically d -variate normal, and asymptotically efficient, which can be summarized as follows:

$$\sqrt{N}(\hat{\omega} - \omega) \xrightarrow{D} \text{MVN}(0, \mathcal{F}^{-1}), \quad (63)$$

where the notation $\xrightarrow{D}$ denotes convergence in distribution and $\mathcal{F} = \Delta [\text{diag}(\pi)]^{-1} \Delta$ is the $d \times d$ Fisher information matrix, with the Jacobian

matrix Δ defined as the $C \times d$ matrix of all first-order partial derivatives of the response patterns probabilities with respect to the item parameters, where d is the size of ω :

$$\Delta = \frac{\partial \pi(\omega)}{\partial \omega'}. \quad (64)$$

Given Equation 63, it can be shown that the asymptotic distribution of $(\mathbf{p} - \boldsymbol{\pi})$ is C-variate normal:

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{D} MVN(\mathbf{0}, \Xi), \quad (65)$$

where $\Xi = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ is the multinomial covariance matrix associated with the multinomial distribution (Equation 61). Applying asymptotic distribution theory once again (see, e.g., Cai et al., 2006), the multinomial residual vector $(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is asymptotically C-variate normal:

$$\sqrt{N}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} MVN\left(\mathbf{0}, \Xi - \Delta \mathcal{F}^{-1} \Delta'\right). \quad (66)$$

Lower Order Marginal Probabilities and Reduction Operators

From Equation 66, one of the key insights is that the multinomial residuals themselves may be combined with linear reduction operator matrices to obtain model-implied marginal (or other reduced) probability summaries that can be tailored to address specific model fit evaluation questions in educational assessment research. We will explain these reduction operators with several examples.

Consider the three-item example from Equation 62. There are three first-order marginal probabilities $\dot{\pi}_i$ ($i = 1, \dots, 3$), one per item. There are also three second-order marginal probabilities $\ddot{\pi}_{ij}$ for the unique item pairs ($1 \leq j < i \leq 3$). In general, these probabilities correspond to the n univariate and $n(n-1)/2$ bivariate margins that can be obtained from the full C-dimensional contingency table using a reduction operator matrix (see, e.g., Maydeu-Olivares & Joe, 2005). An example is given here:

$$\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \ddot{\pi}_{21} \\ \ddot{\pi}_{31} \\ \ddot{\pi}_{32} \end{pmatrix} = \mathbf{L} \hat{\boldsymbol{\pi}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix}, \quad (67)$$

where $\mathbf{L}$ is a fixed operator matrix of 0's and 1's that reduces the response pattern probabilities and proportions into marginal probabilities and proportions up to order 2. $\hat{\pi}_2$ is the vector of first- and second-order marginal probabilities. Correspondingly, $\mathbf{p}_2 = \mathbf{L}\mathbf{p}$ is the vector of first- and second-order observed marginal proportions.

The choice of focusing on the lower order margins is not accidental. As the number of observed variables increases, the number of response patterns increases exponentially. For more than a dozen or so dichotomous variables, the contingency table on which the multinomial is defined becomes sparse for any realistic N . This is known as the sparseness problem in discrete multivariate analysis (see, e.g., Bartholomew & Tzamourani, 1999). Compared with the full multinomial, the lower order margins are significantly better filled under moderate sample size and realistically long educational assessments. Test statistics using marginal probabilities can be significantly more powerful (Cai & Hansen, 2013). Estimation of key parameters in popular models such as IRT models also draw information primarily from the first- and second-order margins. In addition, the second-order marginal residuals contain information about conditional independence, which is critical to model building, as discussed earlier.

More general versions of the reduction operator matrices for models with multiple categorical discrete outcomes can be derived similarly (see, e.g., Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2006). Note that $\mathbf{L}$ has full row rank. It implies that the marginal residual vector $(\mathbf{p}_2 - \hat{\pi}_2) = \mathbf{L}(\mathbf{p} - \hat{\pi})$ is a full-rank linear transformation of the multinomial residual vector $(\mathbf{p} - \hat{\pi})$. Therefore, the marginal residual vector $(\mathbf{p}_2 - \hat{\pi}_2)$ is asymptotically normal:

$$\sqrt{N}(\mathbf{p}_2 - \hat{\pi}_2) = \sqrt{N}\mathbf{L}(\mathbf{p} - \hat{\pi}) \xrightarrow{D} MVN(0, \Xi_2 - \Delta_2 \mathcal{F}^{-1} \Delta_2'), \quad (68)$$

where $\Xi_2 = \mathbf{L}\Xi\mathbf{L}'$, and $\Delta_2 = \mathbf{L}\Delta$ is the Jacobian for the marginal probabilities up to order 2. The dimensionality of the normal random variable is equal to the number of first- and second-order marginal residuals. For example, in the case of dichotomous items, the number is $n(n+1)/2$.

Equation 68 has led to a number of mathematically rigorous model fit indices with known distributional properties. These include limited-information test statistics and additional derived indices for item factor models, hierarchical model, and DCMs (e.g., Hansen et al., 2016; Monroe & Cai, 2015), covering a large segment of the models presented in Table 10.1. Rigorous distribution theory for stepdown tests such as W. H. Chen and Thissen's (1997) local dependence (LD) X^2 that focus on diagnosing the sources of model misfit have also been developed (e.g., Edwards et al., 2018; Liu & Maydeu-Olivares, 2013).

In addition to these statistics, other reduction operator matrices may lead to further refined procedures that target latent variable distribution assumptions (e.g., Li & Cai, 2018) and item-level fit (Orlando & Thissen, 2000). Consider a test with n items and C_i categories for item i . There are $S = 1 + \sum_{i=1}^n (C_i - 1)$ summed scores ranging from 0 to $S - 1$. Suppose the observed summed probabilities based on a sample of size N are equal

to $\bar{p}_s$ for $s = 0, \dots, S-1$. Under maximum likelihood estimation of item parameters, the corresponding model-implied summed score probabilities are formally defined as

$$\bar{\pi}_s = \sum_y 1_s(\|\mathbf{y}\|) \hat{\pi}_y, \quad (69)$$

where $\|\mathbf{y}\| = \sum_{i=1}^n y_i$ is the summed score associated with response pattern $\mathbf{y}$ and the indicator function takes a value of 1 if and only if $s = \|\mathbf{y}\|$, that is,

$$1_s(\|\mathbf{y}\|) = \begin{cases} 1, & \text{if } s = \|\mathbf{y}\| \\ 0, & \text{otherwise} \end{cases}. \quad (70)$$

Equation 70 shows that the model-implied probability for summed score s is a sum over all such response pattern probabilities leading to summed score s ; in other words, it may also be obtained by a reduction operator matrix.

Let $\mathbf{S}$ be this reduction operator matrix. It is a matrix of 0's and 1's such that the premultiplication of $\boldsymbol{\pi}$ by $\mathbf{S}$ yields the summed score probabilities. Each row of $\mathbf{S}$ can be understood as a set of binary logical relations. An element in row j of $\mathbf{S}$ is equal to 1 if and only if the corresponding response pattern in $\boldsymbol{\pi}$ leads to summed score $j-1$. In particular, $\mathbf{S}$ has full row rank and the rows of $\mathbf{S}$ are mutually orthogonal.

Let us return to the three-item example mentioned previously. There are four summed scores in this case: 0, 1, 2, and 3. The 4×8 matrix $\mathbf{S}$ (below) relates the summed score probabilities to the original multinomial probabilities:

$$\bar{\boldsymbol{\pi}} = \begin{pmatrix} \bar{\pi}_0 \\ \bar{\pi}_1 \\ \bar{\pi}_2 \\ \bar{\pi}_3 \end{pmatrix} = \mathbf{S} \boldsymbol{\pi} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \pi_{111} \end{pmatrix}, \quad \hat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_0 \\ \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\pi}_3 \end{pmatrix} = \mathbf{S} \hat{\boldsymbol{\pi}}. \quad (71)$$

The observed summed score proportions can be obtained in a similar way by multiplying the reduction operator with the vector of observed proportions for each response pattern:

$$\bar{\mathbf{p}} = \begin{pmatrix} \bar{p}_0 \\ \bar{p}_1 \\ \bar{p}_2 \\ \bar{p}_3 \end{pmatrix} = \mathbf{S} \mathbf{p} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{pmatrix}. \quad (72)$$

The summed score residual vector $\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}$ is asymptotically S-variate normally distributed, assuming maximum likelihood estimation of the parameters in the model:

$$\sqrt{N}(\bar{\mathbf{p}} - \hat{\bar{\boldsymbol{\pi}}}) = \sqrt{N}\mathbf{S}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{D} MVN(0, \bar{\boldsymbol{\Xi}} - \bar{\Delta}\mathcal{F}^{-1}\bar{\Delta}'), \quad (73)$$

where $\bar{\boldsymbol{\Xi}} = \mathbf{S}\boldsymbol{\Xi}\mathbf{S}' = \mathbf{S}(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')\mathbf{S}' = \text{diag}(\bar{\boldsymbol{\pi}}) - \bar{\boldsymbol{\pi}}\bar{\boldsymbol{\pi}}'$, and $\bar{\Delta} = \mathbf{S}\Delta$ can be understood as the Jacobian for the summed score probability vector. Pragmatically, the Lord-Wingersky (Lord & Wingersky, 1984) algorithm, already popular in IRT analysis and test scoring, could be used to compute the model-implied summed score probabilities (see Thissen & Wainer, 2001) or any further decompositions to target item-specific model fit (see Orlando & Thissen, 2000).

Overall Fit Measures

The foregoing technical discussion highlights the fact that the multinomial residual vector and its linearly reduced variants have well-defined distributional properties. Therefore, they can inform the development of overall fit measures that differ from the traditional approach in the analysis of categorical data such as likelihood ratio G^2 and Pearson's X^2 . On the one hand, these traditional statistics, often called full-information statistics, use residuals based on the full response pattern cross-classifications to test the specified model against the general multinomial alternative:

$$G^2 = 2N \sum_y p_y \log \frac{p_y}{\hat{\pi}_y}, \quad X^2 = N \sum_y \frac{(p_y - \hat{\pi}_y)^2}{\hat{\pi}_y}. \quad (74)$$

Under the null hypothesis that the model fits exactly, these two statistics have the same asymptotic reference distribution, which is a central chi-square with degrees of freedom equal to $C - 1 - d$ (Bishop et al., 1975), when N is large relative to C , which is almost never true in the applications of these models to educational assessment data (Bartholomew & Tzamourani, 1999).

On the other hand, limited-information model fit statistics (e.g., Maydeu-Olivares & Joe, 2005) have become arguably one of the most thoroughly discussed and theoretically understood development in model fit evaluation for discrete multivariate data. As the name suggests, limited-information statistics rely on the linear reduction operators discussed earlier to reduce the multinomial containing full information into distinct summary probabilities.

One class of limited-information fit statistics use residuals based on lower order (e.g., first- and second-order) margins of the contingency table (Cai et al., 2006; Joe & Maydeu-Olivares, 2010). The most prominent member is M_2 . Under the assumption that the number of first- and second-order margins is larger than the number of free parameters d and that Δ_2 has full column rank (local identification), M_2 can be written as

$$M_2 = N(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)' \tilde{\Delta}_2 [\tilde{\Delta}_2' \boldsymbol{\Xi}_2 \tilde{\Delta}_2]^{-1} \tilde{\Delta}_2' (\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2), \quad (75)$$

where $\tilde{\Delta}_2$ is an orthogonal complement of Δ_2 such that $\tilde{\Delta}_2' \Delta_2 = \mathbf{0}$. From Equation 68, $(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is asymptotically normal with zero means and covariance matrix $\boldsymbol{\Xi}_2 - \Delta_2 \mathcal{F}^{-1} \Delta_2'$, which implies that the covariance matrix of $\tilde{\Delta}_2' (\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ is $\tilde{\Delta}_2' \boldsymbol{\Xi}_2 \tilde{\Delta}_2$. Thus, M_2 (for dichotomous data) is asymptotically chi-square distributed with degrees of freedom equal to $n(n+1)/2 - d$. Polytomous versions of the statistic have also been proposed (see, e.g., Cai & Hansen, 2013). The well-established distribution theory of M_2 has also facilitated the derivation of the sample size-independent RMSEA (Maydeu-Olivares & Joe, 2014). Other M_2 -inspired fit indices include the TLI for IRT (Cai et al., 2021).

Beyond Frequentist Model Evaluation in Models for Discrete Data

The new overall model fit statistics described above are technically superior to traditional ones in terms of better decision error rates and statistical power. But perhaps the most intriguing prospect that they offer is the possibility of using our now significantly more refined understanding of the relative sensitivity of each of the pieces to modeling assumption violations, (e.g., first-order margins, second-order margins, summed scores, item-specific summed scores) to help construct statistics that can incorporate principles of model evaluation that are not classical frequency based.

Bayesian Posterior Predictive Checks

In educational measurement, Bayesian posterior predictive model checking (Gelman et al., 1996) is a relatively well-understood tool now thanks to important earlier contributions that examined a number of indices (several based on margins and summed scores) that could serve as the so-called discrepancy measures or test quantities (Levy et al., 2009; Sinharay et al., 2006). In Bayesian posterior predictive model checking, one would need to be able to characterize the entire posterior distribution of all the parameters, for example, with random samples drawn from the posterior. For each random parameter value, we apply the discrepancy measure to the observed data set, yielding the “realized” discrepancies. For each random parameter value, a replicate (plausible) data set can also be produced from the likelihood function. A discrepancy measure may be computed on this replicate data set—the “predictive” discrepancy. The process is typically repeated hundreds, if not thousands, of times in a Monte Carlo simulation, after fitting the model. Plots or numerical summaries of the bivariate distribution of the realized versus predictive discrepancy measures can lead to graphical model fit diagnosis or a Bayesian p value that can help guide model choice.

As an example, Levy et al. (2009) studied a number of discrepancy measures for dimensionality analysis in IRT modeling. These discrepancy measures use information

contained in the item-by-item bivariate distribution to help diagnose the lack of conditional independence. If we were interested in other aspects of model fit, we could choose a different set of discrepancy measures. This point highlights the practically most important aspect of the posterior predictive model-checking procedure. Namely, it can be entirely driven by the needs of the researcher in a particular situation, and there is also no requirement to derive a distribution theory for the discrepancy measures involved. The major contingency is that the model must be estimated with Bayesian parameter estimation approaches. More recently, Lee et al. (2016) proposed approximate procedures inspired by Bayesian posterior predictive model checking but used with standard maximum marginal likelihood estimation. They proposed to replace the samples from the posterior with samples from approximations to the posterior. Kuhfeld (2019) showed that the procedure appears to have performance that is comparable to the original Bayesian procedure while offering additional flexibility.

Information Theory–Based Fitting Propensity Analysis

As a further move beyond frequentist and Bayesian model evaluation techniques, we should address the issue of a model's inherent tendency to fit well to a large number of possible data patterns. This is prominent in the critique against standard goodness-of-fit testing (see S. Roberts & Pashler, 2000, 2002). If the researcher's model is able to fit many possible data patterns well, its excellent fit to a particular data set perhaps offers little evidence in support of substantive theory that led to the model in the first place. Preacher (2006) called this phenomenon the *fitting propensity* (FP) of a model. He also demonstrated that FP not only is a function of the number of parameters (as is customarily intuited), but also is directly related to a model's functional form (i.e., how the parameters are configured in the model's equations). Thus, two models with the same number of free parameters may exhibit wildly different FP.

In earlier literature, FP was also called model *complexity*. Myung and Pitt (1997) defined complexity as "the flexibility inherent in a model that enables it to fit diverse patterns of data" (p. 80). This concept is best captured in the information-theoretic discussions around the minimum description length (Grünwald, 2007; Rissanen, 2007) principle. In the sense of information theory, a statistical model encodes the data as its input. Different models encode data differently. Hence, even for models having an identical number of parameters, if the functional forms of the models are different, the models' FP (or complexity) will be different. Importantly, this is not due to the features of a sample; the difference is attributable to the features of the models under consideration.

To investigate complexity, Bonifay and Cai (2017) fitted a series of latent variable models with an identical number of parameters (but different functional forms in terms of the item-level model involved, nature of latent variables, and constraints on parameters) to randomly generated data uniformly distributed over the space of all possible data sets that can be fit by those models. Replicating Preacher's (2006) results for linear covariance structure models, they found that certain models, such as item exploratory factor analysis and item bifactor models, tend to fit random data sets far better than others.

The concept of model complexity is philosophically and practically important for the statistical modelers and should be considered in conjunction with model fit. When a model fits the data set well, and when the model replicates in other independent holdout samples, it could very well be the case that it is because the model is more complex. That is, such a model naturally fits a large portion of possible data sets. In the language of Popperian philosophy of science, models are mathematical embodiments of theories. On the one hand, a less complex model can make a specific prediction that is more likely falsified in the empirical setting. A more complex model, on the other hand, while doing a good job explaining the observed variations, is less likely to be falsified. As per Occam's razor, one should generally prefer the more falsifiable model.

Concluding Thoughts on Model Fit

Since the publication of the last edition of *Educational Measurement*, substantial progress has been made on the topic of model fit evaluation. We now have access to more accurate limited-information model fit statistics, useful and practical fit indices such as RMSEA and TLI that borrow from concepts developed out of structural equation modeling, and Bayesian posterior predictive model fit checking procedures. We also know, through FP analyses reported in the literature, that some models may have inherently better FP than other models, all else being equal. This suggests that in empirical data analysis measurement researchers should hopefully use a range of possible fit measures that can provide information about different aspects of model fit. The purpose of the analysis and the use of the model's output should be factored into the choices of models. Whenever possible, researchers should not rely on a single model because alternative models may offer additional explanations.

EMPIRICAL EXAMPLE 1

In the first empirical example, we analyze a data set collected from a large-scale randomized, controlled, multisite trial of a math intervention with pre and post measures to illustrate issues around model building, model selection, and fit appraisal. We begin with simpler methods, such as sum scoring and descriptive statistics, followed by applications of more modern methods that make effective use of item-level data and its structure. The goal of our analyses is to illustrate the modeling options and approaches we discussed earlier. We also hope that by providing a more detailed report of our model-building process than typical, the readers could appreciate the decisions involved in statistical modeling of assessment data with realistic degrees of design complexity.

Data Source

The data come from a multisite trial containing item responses collected from 1,664 individuals in 31 treatment classrooms and 29 control classrooms in 23 schools, along with information on school membership and random assignment status (Chung et al., 2014). The original intent was to determine each classroom's assignment (treatment vs. control) randomly within each school. This was complicated by a lack of

consistent implementation or compliance across sites. Some schools would prefer all the classrooms to be in the treatment condition. Many schools have only a single pair of treatment and control classrooms. As such, classroom-level differences are not the main focus of this analysis. Instead, we see the design offering a substantial amount of information regarding treatment effect variations across sites, in addition to the overall statement of the average treatment effect. The control and treatment conditions have 813 and 851 individuals, respectively. The item responses are based on 10 dichotomously scored (0 or 1) items targeting concepts of fractions, such as identifying fractions in the context of a number line and adding fractions. All items are administered at both pretest and posttest. Of the 10 items, 1 is a multiple-choice item with four response options, and the other 9 are fill-in-the-blank items. Figure 10.1 shows 7 of the 10 items used.

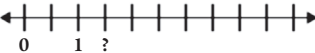
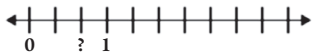
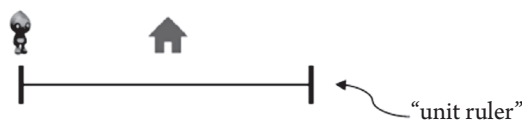
Item description	Example(s)
Adding two fractions (the fraction does not need to be simplified)	$\frac{2}{7} + \frac{1}{7} = \frac{\boxed{}}{\boxed{}}$ $\frac{2}{5} + \frac{3}{10} = \frac{\boxed{}}{\boxed{}}$
At what number is the “?” located?	 
	How far is the puppet from its home? 
Fractional parts of a whole unit	The shaded part of the block below shows $\frac{1}{3}$ of a whole unit. Mark where $\frac{1}{6}$ of a whole unit is. 
	What does the bottom number (4) tell you in $\frac{3}{4}$?
Understand the denominator	a. It tells you there are four fourths in this fraction b. It tells you the whole unit is broken into four pieces c. It tells you there are four whole units in this fraction d. It tells you to add 3 four times

FIGURE 10.1
Example Items Used in Analyses

Summed Score Approach

Summed scoring is the most widely used method for scoring outcome measures, given its simplicity and connection to the classical summed score–based test theory in the social and behavioral sciences (Bauer & Curran, 2015; Curran et al., 2008). Albeit a straightforward method, summed scoring has several limitations when used in randomized multisite studies. The summed score disregards differences in item characteristics (e.g., item difficulty and discrimination), as if each item has the same relationship to the construct being measured. Multiple summed scores become less comparable when the number of items differs between tests or time points because of the study design or missing responses. Past research has also noted other issues and nuances of sum scoring with respect to scoring and scale validity (McNeish & Wolf, 2020) and the multidimensional nature of longitudinal item response data (Cai, 2010a; Cai, Choi, et al., 2016).

Despite these problems, we begin with summed scoring as a baseline for pedagogical purposes. Table 10.2 shows the means and standard deviations of the summed scores of the outcome measure by treatment assignment conditions and occasion (pretest or posttest). The treatment groups are similar at pretest, in terms of both the group means and their standard deviations. There is also a noticeable difference in the change from the pretest to posttest means across conditions. Also as expected, we notice that the variability of the outcome measure increased slightly from pretest to posttest, perhaps more so for the treatment condition.

To break things down further, we can visualize the means and standard deviations of pretest and posttest summed scores by schools and conditions (Figure 10.2). For schools with both control and treatment conditions, most had between-condition differences at pretest falling within one point. There was also variability across schools in mean scores and pre–post changes.

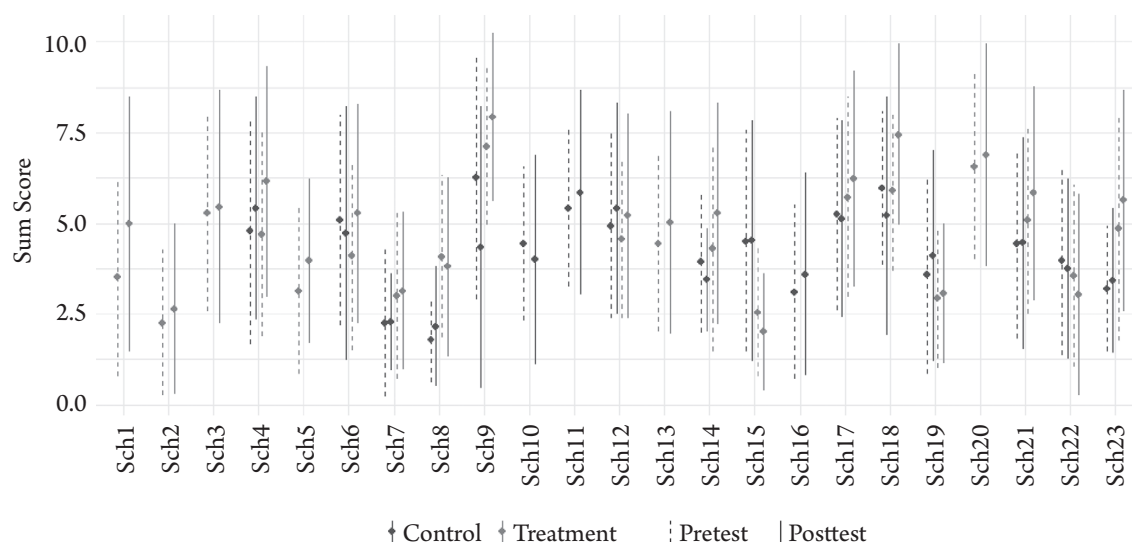
To account for such variations across schools, one approach is to employ the traditional HLM (Raudenbush & Bryk, 2002), which is widely used to estimate the treatment effect in multisite evaluation studies. With a standard two-level model, we specify the Level 1 (within-site) and Level 2 (between-site) equations as follows:

$$y_{ij} = \beta_{0j} + \beta_{1j} * Trt_{ij} + \beta_{2j} * Pre_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2)$$

Table 10.2 Descriptive Statistics of Sum Scores of Pretest and Posttest by Condition

Condition	Time	<i>M</i>	<i>SD</i>
Control (<i>n</i> = 813)	Pretest	4.36	2.73
	Posttest	4.42	2.96
Treatment (<i>n</i> = 851)	Pretest	4.39	2.74
	Posttest	5.02	3.13

Note. Minimum possible value is 0; maximum possible value is 10.

**FIGURE 10.2****Descriptive Statistics of Pretest and Posttest Sum Scores by Schools and Conditions**

Note. The points indicate means and the lines cover the interval of [mean—standard deviation, mean + standard deviation]

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j}, u_{0j} \sim N(0, \tau_{00}) \\ \beta_{1j} &= \gamma_{10} + u_{1j}, u_{1j} \sim N(0, \tau_{11}) \\ \beta_{2j} &= \gamma_{20} + u_{2j}, u_{2j} \sim N(0, \tau_{22})\end{aligned}$$

Here, y_{ij} denotes the outcome variable, that is, the observed posttest summed score from individual i in school j . The pretest summed score Pre_{ij} is school-mean centered and included as a covariate. The treatment indicator variable Trt_{ij} takes a value of -0.5 for the control condition or 0.5 for the treatment condition within each school j . This coding scheme allows β_{0j} to represent the mean posttest summed score for school j and β_{1j} to represent the expected difference in posttest summed score between the two conditions in school j after holding pretest constant.

Table 10.3 presents the HLM results. After controlling for pretest differences, the overall treatment impact ($\hat{\gamma}_{10}$) is 0.59 ($SE = 0.24$). Given that the pooled standard deviation of posttest is roughly 3.08 , we can convert the unstandardized regression coefficient into a standardized estimated effect size (Cohen's $d = 0.59/3.08 = 0.19$) to facilitate comparisons with later analyses.

Item Analysis With Pretest Items

We now turn to item analysis, an important step leading up to the analysis of outcome measures with more refined methods. One key purpose of item analysis is to identify the structure of the item response data—that is, dimensionality evaluation—through which we also address issues of model building, model selection, and fit appraisal. For example, we may want to know whether a set of items represents a single construct or multiple nontrivial ones. In the context of an IRT analysis, we may be concerned with

Table 10.3 Hierarchical Linear Model Results Based on Test Sum Scores

	Model 1: Unconditional			Model 2: Pretest as Covariate		
Fixed Effects	Estimate	SE	p	Estimate	SE	p
Intercept (γ_{00})	4.63	0.24	< .001	4.68	0.23	< .001
Treatment (γ_{10})	0.70	0.33	.036	0.59	0.24	.013
Pretest (γ_{20})				0.70	0.03	< .001
Variance components						
σ^2	2.85			2.21		
τ_{00}	1.05			1.10		
τ_{11}	1.52			0.77		
τ_{22}				0.01		
Log-likelihood	−4142.17			−3722.85		
Akaike information criterion	8296.34			7465.69		
Bayesian information criterion	8328.84			7519.86		

Note. All fixed regression coefficients and the variance components are estimated directly using maximum likelihood.

whether a unidimensional IRT model provides acceptable fit to the item data or if a MIRT model should be used instead. In addition to dimensionality assessment, item analysis provides information on the quality and performance of test items and adds validity evidence to score interpretations.

In what follows, we conduct item analysis on the students' responses to 10 items at pretest with progressively more complex methods discussed in earlier sections, ranging from classical summed score-based statistics, to IRT models, to bifactor models, to DCMs. We use the pretest item responses to show key components of the item analysis before addressing the longitudinal aspects of the example. As shown in Figure 10.1, items used in this example appear to address multiple subskills or concepts within the broader concept of fractions knowledge. However, we will start with unidimensional approaches to underscore the consequences of overlooking design insights and dimensionality and to explore methods for evaluating dimensionality. All models demonstrated for item analysis are estimated in flexMIRT 3.65 (Cai, 2022) using the Bock–Aikin EM algorithm with standard errors (SEs) estimated via the Richardson extrapolation method (e.g., Jamshidian & Jennrich, 2000).

Classical Item Statistics

Table 10.4 presents the classical item statistics—means and standard deviations of item score, correlations between item score and total score, and coefficient

Table 10.4 Classical Item Statistics

Item	Item Score		Item-Total Correlation	Alpha if Deleted
	<i>M</i>	<i>SD</i>		
1	.84	.36	.33	.78
2	.50	.50	.63	.74
3	.48	.50	.61	.74
4	.26	.44	.50	.76
5	.29	.45	.33	.78
6	.24	.42	.48	.76
7	.56	.50	.56	.75
8	.19	.39	.41	.77
9	.66	.47	.43	.77
10	.58	.49	.26	.79

Note. The overall coefficient alpha is .78.

alphas—computed for the listwise-complete data ($n = 1,473$). In general, it is useful to conduct and document these summaries. They often reveal features of the data that should be considered in more complex modeling that is to follow. Because all items are dichotomously scored, the mean value (M) is the proportion of individuals that correctly answered the item and hence is tied to the difficulty of the item. In this case, the hardest and the easiest items appear to be Item 8 ($M = 0.19$) and Item 1 ($M = 0.84$), respectively. The point-biserial item-total correlation compares an item score to the total score from an individual. The size of an item-total correlation provides information on how well an item differentiates between high-performing and low-performing individuals. For example, Item 2 has the highest correlation (.63), while Item 10 has the lowest correlation (.26). Together with the estimates of difficulty, the summaries suggest that the items have a reasonable chance of surviving further scrutiny under more complex models such as IRT.

Unidimensional IRT Model

Moving to the IRT paradigm, we start with a unidimensional IRT model that assumes conditional independence of the items on a single latent factor. Note that fitting a unidimensional model to multidimensional data ignores the dependencies among items forming clusters and will result in distorted estimates of factor loadings, or equivalently, item discrimination parameters (Reise et al., 2007; Steinberg & Thissen, 1996). We may also find signs of misfit based on item fit and model fit statistics.

Of the 10 items, the first 9 were modeled using the 2PL model (Equation 7), in which items varied in their intercept and slope (or discrimination). The last item, a multiple-choice item with four response options, was modeled using the 3PL model (Equation 10) with the lower asymptote parameter (or, equivalently, the pseudo-guessing

parameter) to account for the possibility of guessing. A normal prior distribution $N(-1.1, 0.5)$ was imposed on the *logit* of this parameter. The prior mean of -1.1 yields a prior mode around 0.25 for the $(0-1)$ pseudo-guessing probability. The standard deviation of 0.5 allows for a sufficiently diffuse prior, permitting the estimated pseudo-guessing parameter to move away from the prior mean as needed (e.g., Houts & Cai, 2022, p. 21). Last, the mean of the latent factor was fixed at 0 and its variance was fixed at 1. With this specification, there were $9 \times 2 = 18$ item parameters associated with the 2PL model, $1 \times 3 = 3$ item parameters associated with the 3PL model, and a total of $18 + 3 = 21$ model parameters to be estimated.

Table 10.5 presents the item intercept and slope estimates, as well as the converted factor loading parameters for the 10 pretest items using formulae under Equation 6. Table 10.5 shows a clear variation in the estimated item slopes and factor loadings. In particular, the first 4 items, highlighted in gray, had much higher loadings than the remaining 6 items, thereby potentially “stealing” or pulling the latent factor toward the cluster defined by these 4 items, leaving the underlying latent factor a significantly less important predictor of student performance on the remaining 6 items. Examining the content of the items, the first 4 items and the remaining items indeed appear to form two distinct clusters. In such a situation, it may be inappropriate to assume that the assessment data are unidimensional, making results using summed score-based or unidimensional IRT less desirable or informative.

Hopefully by now it is clear why the analyst’s choices should be informed by the design of assessments. Items presented in Figure 10.1 suggest that the items were designed with multiple subskills or concepts in mind. Items 1–4 targeted the concept of adding fractions, Items 5–9 targeted the concept of identifying fractions in the context of a number line, and Item 10 targeted the concept of understanding the meaning of the

Table 10.5 Unidimensional Item Response Theory Model Item Parameters

Item	Intercept (SE)	Slope (SE)	Loading (SE)
1	2.57 (0.19)	1.83 (0.20)	0.73 (0.06)
2	−0.15 (0.28)	8.70 (1.38)	0.98 (0.01)
3	−0.50 (0.20)	5.77 (0.56)	0.96 (0.01)
4	−2.07 (0.19)	2.69 (0.27)	0.84 (0.04)
5	−1.01 (0.07)	0.73 (0.08)	0.39 (0.06)
6	−1.54 (0.09)	1.27 (0.11)	0.60 (0.06)
7	0.29 (0.08)	1.30 (0.11)	0.61 (0.05)
8	−1.82 (0.10)	1.16 (0.11)	0.56 (0.06)
9	0.83 (0.08)	1.03 (0.10)	0.52 (0.06)
10	−0.11 (0.24)	0.66 (0.12)	0.36 (0.10)

Note. SE = standard error.

denominator. Had the item developers designed the items with only one type of skill in mind, such as adding two fractions with equal denominators, dimensionality issues might not be as dramatic as what we observed.

As further illustrations of their potential value, model fit indices also shed light on the presence of item clusters and data multidimensionality. Table 10.6 shows the item-specific marginal chi-square statistics based on the observed versus expected first-order margins and the LD X^2 indices of W. H. Chen and Thissen (1997) that provide pairwise diagnostic information on possible LD and its directionality. Notice that all of these values are based on the reduction operations introduced earlier (see Equation 67 and surrounding definitions). The LD X^2 values have been standardized with the expected chi-square mean and standard deviation for the ease of interpretation. Each of the first-order marginal chi-square values may be tested against a chi-square variable with one degree of freedom to examine item fit at the level of univariate margins. For each item pair, as suggested by Houts and Cai (2022), the standardized LD X^2 indices can be checked against a threshold (e.g., 3.0) for potential LD. The letter “p” is a convenience that standard IRT software tools now routinely offer. It indicates positive LD, where the observed pairwise association is higher than the model-implied association. The letter “n” indicates negative LD, where the observed association is lower than the model-implied association. As seen in Table 10.5, Items 5–10 showed an unusually high amount of LD in the positive direction. This is consistent with our earlier observation: that the current single latent factor cannot explain all the item-by-item associations among Items 5–10.

Before moving on to another model, we used the limited-information fit statistics M_2 to gauge overall fit: $M_2(34) = 1070.94(p < .001)$, $RMSEA = 0.14$. Recall that M_2 statistic (Equation 36) is computed using univariate and bivariate marginal residuals and is asymptotically chi-square distributed under the null hypothesis of exact model fit (Maydeu-Olivares & Joe, 2005). As noted earlier, on the one hand, exact fit may be

Table 10.6 Unidimensional Item Response Theory Model: Marginal X^2 and Standardized Local Dependence X^2 Statistics

Item	Marginal Chi-Square	1	2	3	4	5	6	7	8	9
1	0.0									
2	0.3	5.5p								
3	0.2	6.4p	2.7p							
4	0.0	2.6p	4.1p	−0.1p						
5	0.0	2.3n	0.9n	4.9n	1.8n					
6	0.0	2.3n	5.6n	7.9n	−0.3n	26.4p				
7	0.0	5.8n	9.2n	5.3n	2.0n	20.8p	38.7p			
8	0.0	4.9n	4.2n	4.5n	2.1n	20.6p	31.7p	31.9p		
9	0.0	5.3n	6.0n	3.9n	0.6n	5.4p	15.7p	137.7p	14.4p	
10	0.0	−0.4n	1.1n	2.5n	−0.7n	12.3p	12.4p	12.6p	4.3p	−0.4p

implausibly strict, and under a large sample such as the situation here, the chi-square will be significant even for very small amount of model error. RMSEA, on the other hand, is a fit measure that takes the influence of sample size into account. We rejected the unidimensional model on an absolute basis, and RMSEA also signals poor fit based on findings reported by Monroe and Cai (2015).

The empirical evidence so far does not support the use of a unidimensional IRT model. We now consider the following information prior to applying a MIRT model to address the lack of conditional independence: (a) all items are developed to measure a board common construct, the knowledge of fractions; and (b) Items 5–9 include number lines in their prompts, although their formats and specific knowledge specifications slightly differ. In this case, a bifactor IRT model could provide an appropriate representation of the item response structure. We now proceed to test that hypothesis.

Bifactor Model

Typically, a bifactor model has one *primary factor* representing the target construct and one or more conditionally orthogonal *specific factors*. These specific factors may represent (a) residual dependence not accounted for by the primary factor or (b) subdomains formed by item specifications (e.g., content, format). All items may load on a primary factor and at most one group/specific factor. When some items load only on the primary factor, some authors call the model an incomplete bifactor model (F. F. Chen et al., 2006). As mentioned earlier, Wainer et al.'s (2007) testlet model is formally nested within the bifactor model. This model adds only as many new parameters to the unidimensional model as there are group/specific factors, making it highly compact and more parsimonious than the general bifactor model. These group/specific factor variance components can also be interpreted as random intercepts for the specific items involved.

Recall that Items 1–4 and Items 5–10 form two clusters that might be distinct from each other. With this in mind, we set up an incomplete bifactor model with Items 1–4 loading on one extra factor specific to these 4 items and Items 5–10 loading on the primary factor only. This allows the general dimension to be defined by all the items, while permitting a group/specific factor to absorb residual dependence above and beyond the general factor. For computational stability, because there are only two clusters, the incomplete model can be substantially more stable than a model that has two group/specific factors. There are only two available clusters, so the choice of which specific cluster of items receives the influence of the group factor is a matter of indifference if one considers the variance decomposition of general versus group/specific factors. However, choosing to use the first 4 items to define a specific factor versus using Items 5–10 does make a difference in terms of the meaning of the general factor. In our case, we are removing any effect unique to Items 1–4 from the general dimension.

A 2PL item-level model was used for Items 1–9 (18 estimated item parameters), and a 3PL item-level model was used for Item 10 (3 estimated item parameters). We further placed testlet response theory-inspired constraints on the 2 item slopes for each of Items 1–4, making them equal within the item. For both the primary factor and one specific

factor, the means were fixed at 0. The variance of the primary factor was fixed at 1, while the variance of the specific factor was freely estimated. In total, there were $18 + 3 + 1 = 22$ freely estimated model parameters, one more than the unidimensional model.

With this constrained bifactor model, most of the factor loadings on the primary factor decreased relative to the factor loadings obtained from the unidimensional model. This was especially true for Items 1–4, of which the loadings ranged between 0.44 and 0.54 (cf. loading sizes of 0.73–0.98 in the unidimensional model). The variance of the specific factor was 2.19 ($SE = 0.29$), more than twice the fixed primary dimension variance of 1.0, indicating very strong excess correlation above and beyond the general factor.

Examining the model fit statistics in Table 10.7, -2 times the log-likelihood at convergence differed by hundreds as we transitioned from a unidimensional model to a constrained incomplete bifactor model, estimating one additional parameter (and losing a single degree of freedom). Other model fit statistics (RMSEA, TLI, Akaike information criterion/Bayesian information criterion) also improved in the same direction, strongly indicating the preference of the incomplete bifactor model over the unidimensional model. For example, the TLI is an incremental fit index based on the M_2 limited-information model fit statistics, with higher values generally indicative of better model fit. Cai et al. (2021) tentatively suggested that values above .96 indicate adequate model fit. The observed value is .99. Though not reported here, the standardized LD X^2 statistics also improve tremendously, and they no longer show any pattern of high LD.

Multilevel Bifactor Model

The multisite study data have a nested structure with individuals nested within sites (schools). Building on the results from bifactor models, we can apply a two-level bifactor model to the pretest item data, where a Level 2 random intercept latent factor is added to

Table 10.7 Model Fit Statistics: Unidimensional Item Response Theory Model Versus Bifactor Model

Statistics	Unidimensional IRT Model	Bifactor Model With One Specific Factor
No. parameters	21	22
$-2 \times \text{Log-likelihood}$	16,005.81	15,291.10
Akaike information criterion	16,047.81	15,335.10
Bayesian information criterion	16,161.56	15,454.27
M_2	1,070.94	83.81
Root-mean-square error approximation	0.14	0.03
Tucker–Lewis index	0.79	0.99

Note. Tucker–Lewis (nonnormed) fit index based on M_2 with the zero-factor null model.

account for any potential variation across schools. A total of 23 model parameters were estimated, including those outlined in the single-level bifactor model with one specific factor, and one additional parameter: 21 item parameters (held equal across levels), the variance of the specific factor, and the variance of the Level 2 factor. Results showed that the estimated variance of the Level 2 factor was 0.26. With the Level 1 (within-school) variance set to 1.00, the intraclass correlation is $0.26 / (1.00 + 0.26) = 0.21$, which suggests a nonnegligible amount of between-school variation. Model fit statistics on this two-level model are shown in Table 10.8. Comparing to those obtained from the incomplete bifactor model, the two-level bifactor model offers a further improvement in fit. It is also more consistent with the design of the overall study. We are, however, also aware that Bonifay and Cai (2017) suggested that the bifactor model may have a high degree of fitting propensity. Thus, the enthusiasm generated by the excellent model fit should perhaps be tempered because alternatives may also fit well.

Diagnostic Classification Model

We examine such an alternative model here. We consider a higher order DCM with discrete latent variables, specifically the DINA model. A core component of DCMs is the Q-matrix. If an item i requires a particular attribute k , the (i,k) th entry in the Q-matrix is 1, and 0 otherwise. In our case, there are 10 items and 5 attributes. The Q-matrix in Table 10.9 is specified based on the knowledge specifications (KS) that link each item with one or more of the five specifications listed here. These specifications were created prior to the actual writing of test items, based on an analysis of the content standards around fractions.

- KS1: Identical (common) units can be added to create a single numerical sum.
- KS2: Each whole unit or part of a whole unit (fractions) can be further broken into smaller, identical parts, if necessary.
- KS3: Positive nonintegers are represented by fractional parts of the interval between whole numbers.

Table 10.8 Model Fit Statistics: Single-Level Bifactor Model Versus Two-Level Bifactor Model

Statistics	Bifactor Model With One Specific Factor	Two-Level Bifactor Model
No. parameters	22	23
–2 Log-likelihood	15,290.05	15,149.81
Akaike information criterion	15,334.05	15,195.81
Bayesian information criterion	15,453.22	15,320.40

Note. Both models use the same bifactor representation with one primary factor and one specific factor; the difference lies in whether the model includes a Level 2 (between-school) latent variable.

Table 10.9 Q-Matrix Specified for Diagnostic Classification Models

Item	KS1	KS2	KS3	KS4	KS5
1	1	0	0	0	0
2	1	1	0	0	0
3	1	1	0	0	0
4	0	1	0	0	0
5	0	0	1	0	0
6	0	0	1	0	1
7	0	0	1	1	0
8	0	0	0	1	0
9	0	0	0	1	0
10	0	0	0	0	1

Note. KS = knowledge specification. Entries with 1 indicate that an item is linked to a given attribute.

- KS4: The size of a rational number is relative to how one whole unit is defined.
- KS5: The denominator of a fraction represents the number of identical parts in one whole unit. That is, if we break the one whole unit into “x” pieces, each piece will be “1/x” of the one whole unit.

One feature of DCMs is the classification of individuals’ attribute, knowledge, or skill mastery at a given time. Given five knowledge specifications, there are five latent attributes and $2^5 = 32$ possible mastery versus nonmastery patterns, termed *attribute profiles*. In using a DINA model, we are assuming that individuals must master all the attributes required by an item to answer that item correctly (the “and” gate). This is accomplished by setting the main effects to zero and estimating only the highest—in this case, second-order—attribute interaction terms for the items targeting more than one attribute, resulting in a total of eight possible inputs in the model: five main effects (KS1 to KS5) and three interaction terms (KS1–KS2, KS3–KS4, and KS3–KS5).

In Table 10.10, we present a subset of the estimated proportions of attribute profiles by conditions. Because we fitted the model on the pretest item data, differences in the estimated proportions between conditions would provide us another layer of assurance regarding the success of random assignment. For example, the percentages of individuals mastering all five attributes (Profile 11111) were 16.0% and 15.7% in the control and treatment conditions, respectively. Similarly, the percentages of individuals mastering none of the attributes (Profile 00000) were 4.7% and 6.1% in control and treatment conditions, respectively. Across all profiles, the mean absolute difference in the estimated proportions between the control and treatment conditions was approximately 0.02, which is quite small.

Model fit statistics of this DCM model, however, indicate worse fit than the bifactor model presented above: $M_2(29) = 227.63 (p < .001)$, RMSEA = 0.06. Inspection of item fit statistics showed higher LD between certain pairs of items, particularly for

Table 10.10 Top Five Estimated Proportions of Attribute Profiles in Diagnostic Classification Models

Condition	KS1	KS2	KS3	KS4	KS5	Proportion (Percentage)
Control	0	0	0	0	0	4.7
	1	0	0	0	0	10.2
	1	1	1	0	0	5.0
	1	1	1	1	0	3.3
	1	1	1	1	1	16.0
Treatment	0	0	0	0	0	6.1
	1	0	0	0	0	10.3
	1	0	1	1	0	4.0
	1	1	1	0	0	3.5
	1	1	1	1	1	15.7

Note. KS = knowledge specification. Only profiles that occur at least once are included. 1 = mastered, 0 = not mastered.

those involving Item 8. Based on research by Bonifay and Cai (2017), we know that DCMs have lower fitting propensity than typical multidimensional IRT models. In other words, a good model fit from a DCM gives us stronger evidence in support of the DCM than a similarly good fit would do to the evidentiary argument in support of, say, a bifactor model. But the degradation of model fit is far too substantial to ignore in this case, despite having four more parameters than the incomplete bifactor model. We chose not to explore the DCM option further.

Longitudinal Latent Variable Models

The prior work on item analysis informs us of the proper latent structure to use and allows us to focus on the changes in latent variables and treatment impact on the changes. We discuss applications of three multilevel item factor models with pre–post latent change parametrization. The first two are two-level models differing in how the treatment assignment status is represented in the data set. The third model is a cross-classified IRT model treating individuals and items as the two crossed factors and school-level influence conditioned out directly using school fixed effects. We realize that the third model is perhaps not ideal for the design that could be exploited to examine treatment effect heterogeneity. Our decision to include it here is pedagogical. Regression modeling with fixed effects is quite popular in evaluation studies, and many analyses only focus on the average treatment effect. We hope to show that econometrics-oriented fixed effects modeling could also be grafted onto complex psychometric-oriented cross-classified IRT models. All three models are estimated in flexMIRT

3.65 (Cai, 2022) using the MH–RM algorithm (Cai, 2010b, 2010c) with SEs estimated recursively (e.g., Cai, 2010b). We discuss applications of three multilevel item factor models with pre–post latent change parametrization, akin to historic models described in Bock & Bargmann (1966) as well as Embretson (1991) and McArdle (2009).

Two-Level IRT Growth Model With Treatment and Control Groups in Separate Data Columns

In the first two-level model, we explicitly represented individuals' treatment assignment status by restructuring the data from the conventional “wide” format to a “block-diagonal” format (Table 10.11). In each school, for individuals assigned to the treatment condition, their item responses occupied a block with $2 \times 10 = 20$ columns because there were 10 respective items for pretest and posttest. Similarly, for individuals assigned to the control condition, their item responses occupied another block with 20 columns. Because each row still represented an individual, the two blocks needed to be placed diagonally.

To model pre–post latent changes, we specified four within-school primary factors to represent the pretest and posttest statuses of the treatment and control groups. This is analogous to conditioning the latent outcome variable on the treatment assignment indicator because it allows the means and variances of latent variables to differ between the two experimental conditions. Because the multisite data have a nested structure

Table 10.11 Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Data Structure

Indv ID	Sch ID	Trt Pre 1	...	Trt Pre 10	Trt Pst 1	...	Trt Pst 10	Con Pre 1	...	Con Pre 10	Con Pst 1	...	Con Pst 10
1	1	1	...	0	1	...	0	–9	...	–9	–9	...	–9
2	1	1	...	0	1	...	1	–9	...	–9	–9	...	–9
3	1	0	...	1	1	...	1	–9	...	–9	–9	...	–9
...	...	...	...	...	1	...	0	...	...	...	...	...	...
23	1	1	...	0	0	...	1	–9	...	–9	–9	...	–9
24	1	1	...	1	1	...	1	–9	...	–9	–9	...	–9
25	1	0	...	0	0	...	1	–9	...	–9	–9	...	–9
26	1	–9	...	–9	–9	...	–9	1	...	0	1	...	1
27	1	–9	...	–9	–9	...	–9	0	...	1	1	...	0
28	1	–9	...	–9	–9	...	–9	0	...	1	1	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
48	1	–9	...	–9	–9	...	–9	1	...	1	1	...	1
49	1	–9	...	–9	–9	...	–9	1	...	1	1	...	1
50	1	–9	...	–9	–9	...	–9	0	...	0	1	...	1

Note. Con = control condition; Pre = pretest; Pst = posttest; Trt = treatment condition. Variable Indv ID denotes the individual ID. Variable Sch ID denotes the school ID. Missing values are coded as –9.

with schools as the Level 2 units, we also specified two between-school primary factors to represent school-level variations in initial status and latent deviation from the initial status. Last, we established cross-factor constraints and freed the means and variances of the latent variables of interest. Table 10.12 presents the factor pattern matrix for such a model. The size of the full factor pattern matrix is 40 (10 items $\times$ 2 time points $\times$ 2 experimental conditions) by 26 (2 between-school primary factors, 4 within-school primary factors, and 20 within-school residual dependence dimensions).¹

Using conventional path diagrams, Figures 10.3 and 10.4 depict the core structures of this model at the between-site and within-site levels. The model is fitted to all 10 items at pretest and posttest, but only four pairs of items are shown because of space constraints. The rectangles represent items, and circles represent latent variables. The items load on two between-school (Level 2) primary dimensions, two within-school

Table 10.12 Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Factor Pattern

Item ID	Between School		Within School											
	Pre	Pst	Con pre	Con pst	Trt pre	Trt pst	Con item 1	Con item 2	...	Con item 10	Trt item 1	Trt item 2	...	Trt item 10
pre_item1	b_1		b_1				b_{1c}							
pre_item2	b_2		b_2					b_{2c}						
...	...		...						...					
pre_item10	b_{10}		b_{10}							b_{10c}				
pst_item1	b_1	b_1	b_1	b_1			b_{1c}							
pst_item2	b_2	b_2	b_2	b_2				b_{2c}						
...	...	...	...	...					...					
pst_item10	b_{10}	b_{10}	b_{10}	b_{10}						b_{10c}				
pre_item1	b_1				b_1						b_{1t}			
pre_item2	b_2				b_2							b_{2t}		
...	...				...								...	
pre_item10	b_{10}				b_{10}									b_{10t}
pst_item1	b_1	b_1			b_1	b_1					b_{1t}			
pst_item2	b_2	b_2			b_2	b_2						b_{2t}		
...	...	...			...	...							...	
pst_item10	b_{10}	b_{10}			b_{10}	b_{10}								b_{10t}

Note. Con = control condition; Pre = pretest; Pst = posttest; Trt = treatment condition. Nonempty entries indicate free parameters.

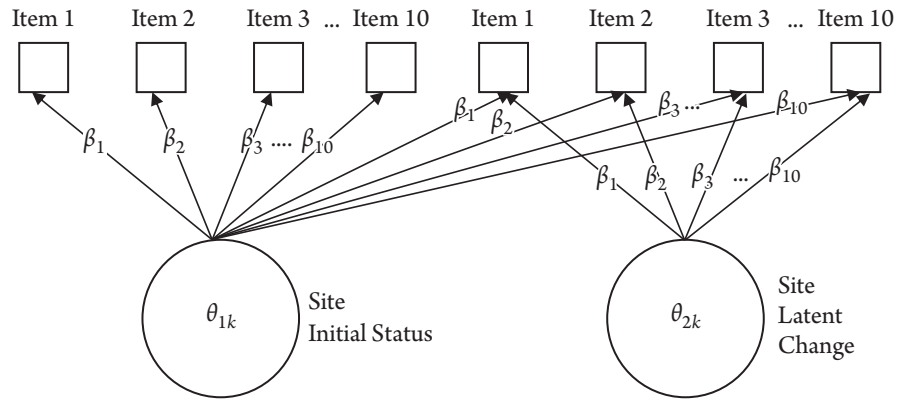


FIGURE 10.3
A Path Diagram Depicting the Level 2 (Site-Level) Model

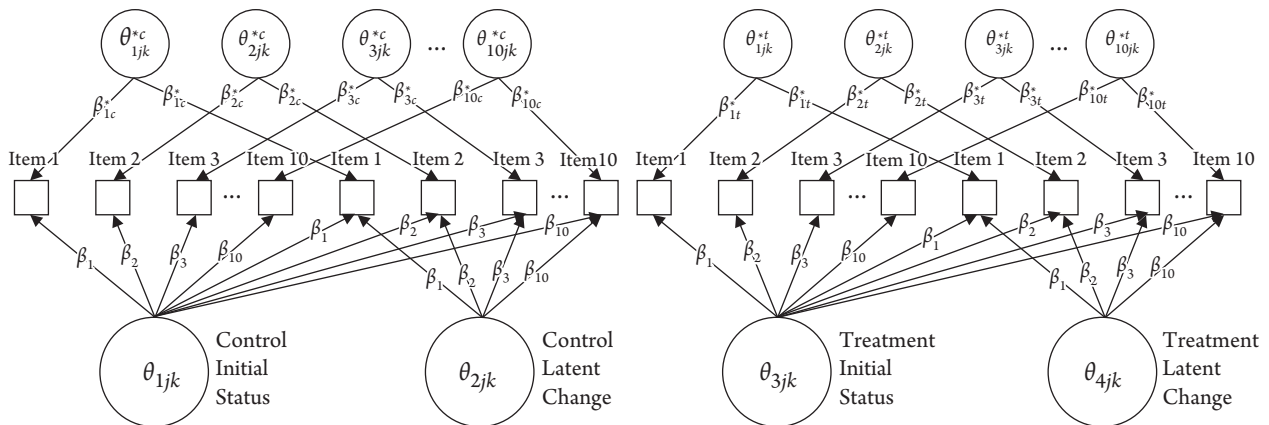


FIGURE 10.4
A Path Diagram Depicting the Level 1 (Individual-Level) Model

(Level 1) primary dimensions, and 10 group-specific (Level 1) residual dependence dimensions in each condition.

Table 10.13 presents descriptive statistics of latent pretest scores and latent change scores by conditions. The means of latent pretest score were -0.02 and -0.46 for the control and treatment conditions, respectively, and their standard deviations were close to 1. The means of latent change scores for the control and treatment conditions were 0.09 and 0.39 , respectively, yielding a difference of 0.30 . The latent change factor could be seen as a difference between the latent pretest factor and the latent posttest factor. Based on variance algebra, because it results from subtraction, the latent change factor typically had a smaller variance or standard deviation relative to the latent pretest factor. This observation holds true for both experimental conditions in this case.

Two-Level IRT Growth Model With Treatment Assignment Indicator as a Predictor

An alternative way to represent the treatment/control conditions in a two-level IRT growth model is to include the treatment assignment indicator as a predictor. In this situation, the data set has the conventional “wide” format, where each row contains the condition indicator, as well as pretest and posttest item responses (Table 10.14).

For this model, we specified one between-school primary factor, to account for variability across schools, and two within-school primary factors to represent the individual-level initial status (pretest) and change. The individual-level pretest latent variable also represents individuals’ deviations from the school random intercept; the latter represents schools’ pretest deviations from the grand mean across all schools (Cai & Houts, 2021). To obtain the estimated treatment effect, we regressed the Level 1 latent variables on the condition indicator.

Table 10.13 Descriptive Statistics of Latent Pretest Score and Latent Change Score by Condition

Condition	Latent Variable	<i>M</i>	<i>SD</i>	Min.	Max.
Control (<i>n</i> = 813)	Pretest	−0.02	0.91	−2.07	2.42
	Change	0.09	0.40	−1.04	1.57
Treatment (<i>n</i> = 851)	Pretest	−0.46	0.99	−2.55	2.20
	Change	0.39	0.52	−0.98	2.04

Table 10.14 Two-Level Item Response Theory Growth Model With Condition Indicator as a Predictor: Data Structure

Individual ID	School ID	Condition	Pre 1	...	Pre 10	Pst 1	...	Pst 10
1	1	1	0	...	0	0	...	1
2	1	1	0	...	0	1	...	1
...	...	...	...	...	...	...	...	...
24	1	1	0	...	0	1	...	1
25	1	1	1	...	1	1	...	1
26	1	0	1	...	1	1	...	1
27	1	0	0	...	1	0	...	1
...	...	...	...	...	...	...	...	...
49	1	0	1	...	0	1	...	0
50	1	0	0	...	0	0	...	1

Note. For the Condition indicator, 1 represents the treatment condition and 0 represents the control condition.
Pre = pretest; Pst = posttest.

Table 10.15 presents key parameter estimates for the primary factors. Results suggested that, on average, the control and treatment groups did not differ significantly at pretest ($\hat{\gamma} = 0.04$, $SE = 0.04$). Also, there was statistically significant variability of pretest across schools (variance of the between-school primary factor = 0.30, $SE = 0.09$). With the individual-level pretest variance constrained to 1, we can compute the intraclass correlation: $0.30 / (0.30 + 1) = 0.23$. Moreover, there was a significant pre-post improvement in the control condition (latent change $M = 0.17$, $SE = 0.02$), reflecting ongoing learning, along with a significant treatment effect ($\hat{\gamma} = 0.31$, $SE = 0.03$) that is the regression coefficient associated with the within-school latent change variable. One may view this model as a refined HLM analysis, replacing summed scores with latent change as the outcome. Relative to the variability of the change dimension ($SD = \sqrt{0.38} = 0.62$), the magnitude of the treatment effect is roughly half a standard deviation.

Cross-Classified IRT Model With Individuals and Items as Crossed Factors

The last model is a cross-classified IRT model that treats the characteristics of items, such as their difficulty, as random parameters, and models individuals and items as two crossed factors. For this model, our data set is structured by converting item responses from each individual into a diagonal matrix, where each row corresponds to an individual-item combination. Table 10.16 presents the example data structure for two individuals belonging to two schools. We again include the condition indicator as a predictor. Because individuals are nested within schools, the school membership is denoted by a set of dummy indicators that take on a value of 1 if the corresponding individual is in the school denoted by the dummy and 0 otherwise. In this situation, school effects are estimated as fixed effects. With this particular data structure, individuals and items are the two crossed Level 2 factors. The Level 1 units are the specific item responses.

Table 10.15 Two-Level Item Response Theory Growth Model With Condition Indicator as a Predictor: Key Results

Parameter	Between-School	Within-School Initial Status (Pretest)	Within-School Change
Latent mean (SE)	0.00 (—)	0.00 (—)	0.17 (0.02)
Regression coefficient (SE)		0.04 (0.04)	0.31 (0.03)
Covariance matrix (SE)	0.30 (0.09)		
	0.00 (—)	1.00 (—)	
	0.00 (—)	0.00 (—)	0.38 (0.02)

Note. Results for within-school residual dependence dimensions are excluded, given that all means are 0 and variances are fixed to 1.

We specified this model as follows. On the individual (or based on syntax in flex-MIRT, cluster) side, we introduced four factors: Two factors represented individuals' initial status at pretest and latent change at posttest, and the other two factors represented the dimension formed by Items 1–4 (as identified in bifactor model analysis) at pretest and posttest. On the item (or block) side, we included one factor representing the relative difficulty of the items. We then constrained the slopes and intercepts to be equal across items loading on the same factor. To estimate the treatment effect, we again regressed the latent pretest and change factors on the condition indicator. To account for school effects, we further regressed the latent pretest factor on 23 school dummy indicators.

Table 10.17 shows the treatment effect estimates. Based on the estimates, the control and treatment groups did not differ significantly at pretest ($\hat{\beta} = 0.003$, $SE = 0.093$). Examining the coefficient relating to the latent change, there was a statistically

Table 10.16 Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Data Structure

Indv ID	Item ID	Cond	Pre 1	Pre 2	...	Pre 9	Pre 10	Pst 1	Pst 2	...	Pst 9	Pst 10	Sch 1	Sch 2	...
1	1	1	0	−9	...	−9	−9	0	−9	...	−9	−9	1	0	...
1	2	1	−9	0	...	−9	−9	−9	0	...	−9	−9	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	8	1	−9	−9	...	−9	−9	−9	−9	...	−9	−9	1	0	...
1	9	1	−9	−9	...	0	−9	−9	−9	...	1	−9	1	0	...
1	10	1	−9	−9	...	−9	0	−9	−9	...	−9	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
51	1	1	1	−9	...	−9	−9	1	−9	...	−9	−9	0	1	...
51	2	1	−9	0	...	−9	−9	−9	0	...	−9	−9	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
51	8	1	−9	−9	...	−9	−9	−9	−9	...	−9	−9	0	1	...
51	9	1	−9	−9	...	0	−9	−9	−9	...	0	−9	0	1	...
51	10	1	−9	−9	...	−9	0	−9	−9	...	−9	1	0	1	...

Note. Variable Cond denotes the condition, where 1 denotes the treatment condition and 0 denotes the control condition. Variable Indv ID denotes the individual ID. Variables with the prefix Sch are the school membership dummy indicators. Missing values are coded as −9. Pre = pretest; Pst = posttest.

Table 10.17 Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Key Results

Regression Coefficient	Estimate	SE
Initial Status On Treatment Indicator	0.003	0.093
Latent Change On Treatment Indicator	0.982	0.197

Note. Estimates of school effects are not shown here because of space constraints.

significant and positive treatment effect on posttest deviations from the initial status (i.e., improvements in performance) ($\hat{\beta} = 0.982$, $SE = 0.197$). Estimates of school fixed effects ranged from -0.97 (School 2) to 1.59 (School 20).

Summary of Empirical Example 1

In sum, we have conducted classical and IRT-based item analysis. We noticed that the results appear to confirm the presence of additional dimensions due to assessment design. We explored several plausible models, including the item bifactor model, the multilevel bifactor model, and a DCM. We used a range of fit measures to help guide the down selection. We also incorporated the IRT model directly into several multilevel or cross-classified latent variable models that represented the school effects, random assignment, and the pre–post design. These models further refine the traditional HLM analyses. These models are all suggestive of the presence of a significant and practically noticeable treatment effect.

EMPIRICAL EXAMPLE 2

In this second empirical example, we examine a data set collected from an operational field test of an alternate English language proficiency assessment (*Alt ELPA*) to further illustrate issues around the treatment of item parameters as random effects. The Alt ELPA test is specifically developed for English learners who have significant cognitive disabilities.

Data Source

The field-test data contain item responses collected from 712 students in Grades 2–3 across nine states. The item responses are based on 80 items in two modalities: 40 items assessing receptive modality (listening or reading domains) and 40 items assessing productive modality (speaking or writing domains). All 40 receptive items and 28 productive items are dichotomously scored, and 12 productive items are polytomous scored (0, 1, 2, or 3). Students were randomly assigned forms (A or B) for the domains they were not exempted from and were presented with the items that corresponded to the assigned forms. As a result, students were presented with 40 items, if not exempted from any domain, or fewer if exempted from one or more of the domains. For example, consider a scenario in which a student is exempted from writing and is assigned forms A, B, and B for the remaining domains. In this case, the student will be administered 30 items from reading form A, listening form B, and speaking form B. Table 10.18 provides a summary of the characteristics of the forms. The main challenges are that the sample size is relatively small, and the data contain substantial missingness due to domain exemptions resulting from disabilities. These challenges are not due to sampling; the target population itself is small and heterogeneous, including dramatic variations in the type of cognitive disabilities.

Table 10.18 Summary of Alt ELPA Grade 2–3 Item Characteristics

Modality	Domain	Form A		Form B	
		Dichotomous	Polytomous	Dichotomous	Polytomous
Receptive	Reading	10	—	10	—
	Listening	10	—	10	—
Productive	Speaking	7	3	7	3
	Writing	7	3	7	3

Multidimensional IRT Model

The assessment designers and educators involved in the project envisioned a representation of English language proficiency composed of two modalities: receptive and productive. This corresponds nicely with a two-dimensional item factor model. The two correlated dimensions corresponded to the two modalities. The correlation can also serve to produce modality scaled scores that can borrow strength from the other modality (for more discussions on subscore augmentation, see Thissen & Wainer, 2001).

The fitted item factor model had an independent cluster structure (McDonald, 2000), where each item loads on only one of the two modalities. At the item level, the correct response probabilities of dichotomous items were modeled using the 2PL model (Equation 7), and those of polytomous items were modeled using the GPC model (Equation 9). To stabilize estimation under the limited sample size and considerable missingness by design, we constrained the item slope parameters to be equal within the groups of receptive dichotomous (RD) items, productive dichotomous (PD) items, and productive polytomous (PP) items. The next to last column of Table 10.19 shows the summary and key results of the multidimensional IRT model.

Cross-Classified IRT Model

The multidimensional model above treats item parameters as fixed effects and person parameters as random effects. Alternatively, one could treat the item responses to be cross-classified by items and persons, making both items and persons random facets. An immediate computational benefit of the alternate specification is that the number of structural parameters dramatically decreases, which stabilizes parameter estimation under small sample size.

We specified the model as follows. On the individual side, we included two correlated factors corresponding to individuals' proficiency in receptive and productive modalities, which we term person effects. On the item side, we introduced five factors: One factor indicates RD items' deviation from their overall difficulty, another factor indicates PD items' deviation from their overall difficulty, and the other three factors indicate PP items' deviations from their overall category-specific difficulties; we term these factors item effects. Each of the item effects was assumed to follow a normal distribution with mean zero and freely estimated variance. As in the standard

multidimensional IRT model, item slope parameters were constrained to equality within RD, PD, and PP items, and the slope parameters were considered fixed. The person scores were computed by a multiple imputation-based procedure (Huang et al., 2024; Yang et al., 2012).

For this model, the data set is restructured into a long format by collapsing the 80 items into three pseudo-items corresponding to the three item types—RD, PD, and PP—that are assumed to have equal item slopes. Table 10.20 presents the example data structure for two individuals.

The last column of Table 10.19 shows the summary and key results of the cross-classified IRT model. The point estimates closely matched those found in the MIRT model. The correlation between item difficulty estimates in the two models was .998 for dichotomous items and .997 for polytomous items. Similarly, the correlation between person scores from the two models was .994 for receptive modality and .991 for productive modality. In other words, the cross-classified model is entirely capable

Table 10.19 Model Fitting Summary and Key Results: Multidimensional Item Response Theory (IRT) Model Versus Cross-Classified IRT Model

		Multidimensional IRT Model	Cross-Classified IRT Model
Parameters	Person Scores	Random	Random
	Item Intercept	Fixed	Random
Number of latent variables (dimensionality)		2	7
Number of parameters		108 (1 correlation between person effects + 3 slopes for person effects + 68 intercepts for dichotomous items + 12 × 3 intercepts for polytomous items)	14 (1 correlation between person effects + 3 slopes for person effects + 5 variances of item effects + 5 overall intercepts [1 for RD, 1 for PD, and 3 for PP])
Estimates ^a	Correlation between person effects	0.822 (0.022)	0.822 (0.013)
	Slope for person effects	RD: 1.072 (0.050) PD: 1.422 (0.081) PP: 2.139 (0.160)	RD: 1.065 (0.032) PD: 1.413 (0.045) PP: 2.112 (0.079)
	Variance of item effects		RD: 0.420 (0.001) PD: 0.578 (0.002) PP: 0.159 (0.002), 0.229 (0.005), and 0.066 (0.002) for polytomous generalized partial credit intercepts

Note. PD = productive dichotomous; PP = productive polytomous; RD = receptive dichotomous.

^a The numbers in the parentheses are standard errors corresponding to the point estimates.

Table 10.20 Cross-Classified Item Response Theory Model: Data Structure (Example 2)

Indv ID	Item ID	Resp RD	Resp PD	Resp PP
1	RD_01	0	−9	−9
⋮	⋮	⋮	⋮	⋮
1	RD_40	1	−9	−9
1	PD_01	−9	1	−9
⋮	⋮	⋮	⋮	⋮
1	PD_28	−9	0	−9
1	PP_01	−9	−9	3
⋮	⋮	⋮	⋮	⋮
1	PP_12	−9	−9	1
2	RD_01	1	−9	−9
⋮	⋮	⋮	⋮	⋮
2	RD_40	0	−9	−9
2	PD_01	−9	0	−9
⋮	⋮	⋮	⋮	⋮
2	PD_28	−9	1	−9
2	PP_01	−9	−9	0
⋮	⋮	⋮	⋮	⋮
2	PP_12	−9	−9	2

Note. Variable Indv ID denotes the individual ID. Variable Resp RD is an item response if the item is receptive and dichotomous and missing otherwise. Variable Resp PD is an item response if the item is productive and dichotomous and missing otherwise. Variable Resp PP is an item response if the item is productive and polytomous and missing otherwise. Missing values are coded as −9. PD = productive dichotomous; PP = productive polytomous; RD = receptive dichotomous.

of offering what the standard fixed-item approach always offered in terms of item parameters and scaled scores.

The cross-classified IRT model yielded substantially smaller standard errors and can produce additional summaries not available in the standard approach. For instance, the correlation between the two person effects were estimated to be .822 in both models, but its standard error was 0.022 in the MIRT model and only 0.013 in the cross-classified IRT model. The smaller standard errors in item parameters imply less subsequent decision error in scaled score estimates and classification decisions, which rely on item parameter estimates.

By specifying both item and person variances, we can directly decompose variability into relevant sources in a similar manner as in generalizability theory. For the 20 RD items in the each of the test forms, for example, 73% of the total explained variance (i.e., $1.065^2 / (1.065^2 + 0.420)$) is attributable to persons, with the rest of the variance attributable to items. For a single RD item with a 2PL IRT model at the item level, the implicit error distribution is that of the standard logistic, with mean 0 and variance equal

to $\frac{\pi^2}{3} \approx 3.29$ (Bauer, 2009). Therefore, a rough index of dependability of absolute decisions with regard to model-based scaled scores in the receptive modality measured by 20 items could be computed as

$$\frac{1.065^2}{1.065^2 + (.42 + 3.29)/20} = .86.$$

This index is similar in spirit to Brennan and Kane's (1977) dependability coefficient in a $p \times i$ design. Without making items a random facet, we would be unable to decompose variances into relevant sources in this nonlinear model because the variance for the item facet would be unavailable.

Cross-Classified IRT Model With a Predictor

Another advantage of treating item parameters as random effects, in addition to the computational and conceptual benefits mentioned in the previous section, is that we can simultaneously study the impact of key predictors on the random effects (Huang & Cai, 2023). In the context of Alt ELPA, we adopted a principled assessment design process with standard setting embedded into the item development process (Lewis & Cook, 2020). In brief, under principled assessment design, each item is intended (hypothesized) to elicit observations that support measurement of a specific achievement level. Thus, the ultimate correspondence between the statistically estimated item difficulty levels and the targeted achievement level (hereafter, target level) is critical in lending support to the embedded standard-setting procedure and the ultimate scaled score interpretation and use. We employed the cross-classified IRT model to integrate the qualitative evidence from the item-writing process with the quantitative evidence from the calibration study using field-test data. We empirically assessed the alignment by incorporating a latent structural model into the measurement model discussed in the previous section, specifically by regressing the items' relative difficulty levels (estimated from the field-test data) on their target levels (assigned during the item-writing process). The target levels had been evaluated in four levels: 1, 2, 3, and 4. The number of items assigned to the four levels was 7 (17.5%), 14 (35%), 13 (32.5%), and 6 (15%), respectively.

For a simple illustration, we selected 40 RD items to apply the same cross-classified IRT model. We specified two latent variables: one person effect representing individuals' proficiency and one item effect representing RD items' relative difficulties. Then, to assess the predictive relationship between the items' target levels and their relative difficulties in the field test, we regressed the item effect on the target levels.

Turning our attention to the regression result, an estimated regression coefficient was $\hat{\beta} = -0.390$. The estimate had a standard error $SE(\hat{\beta}) = 0.043$ and was statistically significant ($p < .001$). The estimated coefficient suggests that a 1-unit increase in the

target level corresponds to a decrease of 0.390 in an item's deviation from the overall RD items' intercept. Since the intercept a product of minus the slope and the difficulty, we can also infer that a 1-unit increase in the target level results in an increase of 0.379 in an item's deviation from the overall RD items' difficulty (obtained by dividing 0.390 by the estimated item slope of 1.028). From analysis of variance perspective, the proportion of item random effect variance (relative item difficulty) explained by the target level predictor was 64%. The regression analysis follows the explanatory tradition of de Boeck and Wilson (2004) that would not be feasible if we had not parameterized and fitted the model as a random item model in the first place. Substantively, these results lend support to the claim that one could indeed produce items that are aligned with their intended target levels. This finding will help refine the assessment development process in the future.

CONCLUSION

A wide range of topics have been covered in this chapter. We started with general discussions around the sources of data and purposes of modeling in educational assessment. We introduced several core statistical modeling concepts, including latent variables, conditional distributions and conditional independence, Bayesian inference fundamentals (likelihood, prior, and posterior), and model fit evaluation in the context of fallible working models that do not fit exactly. We discussed parameter estimation and saw how the missing data formulation could lead to insights that benefit the development of estimation algorithms in statistical computing. We described a general multilevel latent variable modeling framework and connected it to various “named” models that see routine use in practical settings. We reviewed the distribution theory needed for constructing more refined and targeted tests of model fit. Inspired by information theory, we briefly touched on the philosophical dilemma that should be at the forefront of every researcher's mind—that good fit does not automatically lend unequivocal support to a statistical model (or the theory that it encodes), because the model may have an inherent propensity to fit well to many diverse data patterns. We then proceeded to analyze two empirical data sets to help illustrate the points we made earlier. The first example involved an integrated latent variable analysis combining item response modeling with multilevel modeling. The second example illustrated the benefits of using more advanced models with crossed random effects for the test items in the context of language proficiency assessments with special populations. Both examples had a cross-classified modeling component, with the second example further featuring a use case for modeling random item effects and regressing these effects on one explanatory predictor. These models not only offer inference and output comparable to more standard item response models with fixed item effects, but also provide more information about sources of variability. They support statistical modeling that can explicitly incorporate assessment design features in the assessment validation process.

We speculate that several important ongoing trends and developments in educational measurement are likely to impact the practice of statistical modeling. The first is the availability of cheap, distributed, and scalable options to conduct software-based data collection. We hope it is obvious to anyone reading this chapter that educational technology products on the market today or products that we have yet to see may be used for data gathering for assessment purposes. The increasingly voluminous data also prompted the increasing use, perhaps out of necessity, of tools derived from what Breiman (2001) aptly referred to as the algorithmic approach to statistical modeling. The ability to predict outcomes or some criterion in performance is paramount in this approach, rather than the more traditional method that seeks to gain an understanding of what may be part of the mechanisms or processes that led to the data in the first place. These new tools are incredibly powerful and undoubtedly useful. What remains to be seen is whether they will lead to increased ability for us to explain what we observed, or, in other words, more theory building, and conceptual clarity that is afforded by subjecting theories (with models as their embodiments) to statistical tests and rejecting them if necessary.

The second trend is the resurgent interest in connecting the design of assessment systems with the design of learning and instructional systems. In terms of statistical approaches, we must rise to the challenge of combining assessments taken at multiple time points, for example, throughout the school year in K–12 settings, to formulate reliable measurement of progress, achievement, or mastery. We must also be able to use the data more efficiently to provide feedback that is more timely and more relevant to instruction and improvement in learning. We are particularly enthusiastic about the prospects of multidimensional or DCM approaches to help meet this challenge.

The final development is related to the fairness and consequential validity of educational assessments. This is not a new topic. To address fairness issues, it is imperative that modelers of educational assessment data gain a more thorough understanding of the social contexts in which the assessment programs operate (see Ercikan & Solano-Flores, Lane & Marion, Rodriguez & Thurlow, and Zwick, this volume). To arrive at a fair interpretation of the results, one must also understand the historical and societal determinants of observed differences across subpopulations. Finally, the modeler should hopefully seek to understand better the implications of the use of statistical summaries resulting from educational assessments and develop methods to help mitigate any potentially unfair impact on the individuals and to communicate the assumptions and outcomes plainly to the stakeholders.

The first edition of *Educational Measurement* was published more than 70 years ago. Since that time, the theory and practice of statistical modeling have experienced transformative changes. Given the progress, we are hopeful that the readers of this chapter and the researchers in the field will continue to substantially improve the modeling of educational assessment data in the next 70 years.

ACKNOWLEDGMENTS

We would like to thank Tianying (Teanna) Feng and Yun-Kyung Kim for their contributions that greatly improved this chapter. The U.S. Department of Education supported the design and collection of the data sets mentioned in the empirical examples (R305C080015 and S368A150011).

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17, 251–269.
- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.
- Bartholomew, D. J., & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research*, 27, 525–546.
- Bauer, D. J. (2009). A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes. *Psychometrika*, 74(1), 97–105.
- Bauer, D. J., & Curran, P. J. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research* (pp. 3–38). Information Age Publishing.
- Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika*, 48(4), 493–517.
- Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). Routledge.
- Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6), 706–732.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring any examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.
- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Bock, R. D., & Bargmann, R. E. (1966). Analysis of covariance structures. *Psychometrika*, 31, 507–533.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3), 261–280.
- Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35(2), 179–197.

- Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.
- Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, 22(3), 581–596.
- Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, 52(4), 465–484.
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277–289.
- Brillinger, D. R. (2014). *Practice of data analysis: Essays in honor of John W. Tukey*. Princeton University Press.
- Browne, M. W. (1984). Asymptotically distribution-free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62–83.
- Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, 1(2), 103–124.
- Cai, L. (2008). *A Metropolis–Hastings Robbins–Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model* [Unpublished doctoral dissertation]. University of North Carolina at Chapel Hill.
- Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4), 581–612.
- Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307–335.
- Cai, L. (2010c). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, 75(1), 33–57.
- Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, 80(2), 535–559.
- Cai, L. (2022). *flexMIRT® version 3.65: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Vector Psychometric Group.
- Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, 3, 297–321.
- Cai, L., Choi, K., & Kuhfeld, M. (2016). On the role of multilevel item response models in multi-site evaluation studies for serious games. In H. F. O’Neil, E. L. Baker, & R. Perez. (Eds.), *Issues regarding the use of games and simulations for teaching and assessment* (pp. 304–325). Taylor & Francis.
- Cai, L., Chung, S. W., & Lee, T. (2021). Incremental model fit assessment in the case of categorical data: Tucker–Lewis index for item response theory modeling. *Prevention Science*, 24(3), 455–466.

- Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66(2), 245–276.
- Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika*, 86(3), 754–777.
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, 59(1), 173–194.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16(3), 221–248.
- Camilli, G. (1994). Teacher's corner: Origin of the scaling constant $d = 1.7$ in item response theory. *Journal of Educational Statistics*, 19(3), 293–295.
- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of the quality of life. *Multivariate Behavioral Research*, 41(2), 189–225.
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.
- Chen, Y., Moustaki, I., & Zhang, S. (2022). On the estimation of structural equation models with latent variables. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp.145–162). Guilford.
- Choi, H.-J., Rupp, A. A., & Pan, M. (2013). Standardized diagnostic assessment design and analysis: Key ideas from modern measurement theory. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific* (pp. 61–85). Springer.
- Chung, G. K. W. K., Choi, K., Baker, E. L., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.
- Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, 57(3), 357–369.
- Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the “problem” of sample size: A clarification. *Psychological Bulletin*, 109(3), 512–519.
- Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Developmental Psychology*, 44(2), 365–380.
- De Boeck, P. D., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.
- De la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333–353.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

- Dixon-Román, E. J., & Gergen, K. J. (2013). *Epistemology in measurement: Paradigms and practices*. Gordon Commission. https://www.ets.org/Media/Research/pdf/dixon_roman_gergen_epistemology_measurement_paradigms_practices.pdf
- Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497.
- Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23(1), 138–149.
- Embretson, S. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56, 495–515.
- Fanelli, D. (2019). A theory and methodology to quantify knowledge. *Royal Society Open Science*, 6(4), 181055.
- Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Research Report No. 791). University of Birmingham. <http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/cross-classified-models-review-for-dfes.pdf>
- Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–725.
- Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733–807.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3), 423–436.
- Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, 59, 361–368.
- Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, 22(3), 364–375.
- Grünwald, P. D. (2007). *The minimum description length principle*. MIT Press.
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, 69(3), 225–252.
- Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities blending theory with practicality [Unpublished doctoral dissertation]. University of Illinois at Urbana–Champaign.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika*, 57, 97–10.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2), 191–210.
- Ho, A. D. (2014). Variety and drift in the functions and purposes of assessment in K–12 education. *Teachers College Record*, 116(11), 1–18.

- Houts, C. R., & Cai, L. (2022). *flexMIRT® user's manual version 3.6: Flexible multilevel multidimensional item analysis and test scoring*. Vector Psychometric Group.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
- Huang, S. (2021). Estimation of cross-classified multilevel item response theory models with Metropolis–Hastings Robbins–Monro Algorithm [Unpublished doctoral dissertation]. University of California, Los Angeles.
- Huang, S., & Cai, L. (2021). Lord–Wingersky algorithm Version 2.5 with applications. *Psychometrika*, 86(4), 973–993.
- Huang, S., & Cai, L. (2023). Cross-classified item response theory modeling with an application to student evaluation of teaching. *Journal of Educational and Behavioral Statistics*, 49(3), 311–341.
- Huang, S., Chung, S., & Cai, L. (2024). A random item effects generalized partial credit model with a multiple imputation-based scoring procedure. *Quality of Life Research*, 33(3), 637–651.
- Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B*, 62, 257–270.
- Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2017). A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, 82(3), 693–716.
- Jin, I. H., & Jeon, M. (2019). A doubly latent space joint model for local item and person dependence in the analysis of item response data. *Psychometrika*, 84(1), 236–260.
- Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, 75(3), 393–419.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36(3), 347–387.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272.
- Kuhfeld, M. (2019). A posterior predictive model checking method assuming posterior normality for item response theory. *Applied Psychological Measurement*, 43(2), 125–142.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). Wiley.
- Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 206–220.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33(7), 519–537.

- Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39(1), 8–21.
- Li, Z., & Cai, L. (2018). Summed score likelihood-based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, 78(5), 857–886.
- Little, R. J. A., & Rubin D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.
- Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, 73(2), 254–274.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8(4), 453–461.
- MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, 38(1), 113–139.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111(3), 490–504.
- MacCallum, R. C., & Tucker, L. R. (1991). Representing source of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109(3), 502–511.
- Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(3), 333–356.
- Maydeu-Olivares, A., & Joe, H. (2005). Limited and full-information estimation and goodness-of-fit testing in 2ⁿ contingency tables: A unified framework. *Journal of the American Statistical Association*, 100(471), 1009–1020.
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71(4), 713–732.
- Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, 49(4), 305–328.
- McArdle, J. J. (2009). Latent variable modeling of difference and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.
- McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed Models* (2nd ed.). John Wiley & Sons.
- McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24(2), 99–114.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52, 2287–2305.
- McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, 28(1), 61–88.

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131–154.
- Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, 50(6), 569–583.
- Muraki, E. (1992). *A generalized partial credit model: Application of an EM algorithm* (ETS Research Report No. RR-92-6). ETS.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, 4(1), 79–95.
- Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 16(1), 1–32.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146–178.
- Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, 41(3), 227–259.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.
- Reckase, M. D. (2009). *Multidimensional item response theory models*. Springer.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667–696.
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19–31.
- Rissanen, J. (2007). *Information and complexity in statistical modeling* (Vol. 152). Springer.
- Roberts, G. O., & Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4), 351–367.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358–367.
- Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review*, 109(3), 605–607.
- Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22, 400–407.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151–1172.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
- Samejima, F. (1969). *Estimation of ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Psychometric Corporation. <http://www.psychometrika.org/journal/online/MN17.pdf>
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Wiley.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.
- Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, 79(2), 310–334.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298–321.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.
- Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, 1, 81–97.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52(3), 393–408.
- Tal, E. (2020). Measurement in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 ed.). Stanford University. <https://plato.stanford.edu/archives/fall2020/entries/measurement-science/>
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345–354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3), 287–305.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 53–86). Routledge.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567–577.
- Thissen, D., & Wainer, H. (2001). *Test scoring*. Routledge.
- Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics*, 22(4), 425–445.
- Thurstone, L. L. (1935). *The vectors of mind*. University of Chicago Press.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, 86(1), 1–29.

- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287–308.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS No. RR-05-16). ETS.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, 2, 9–36.
- von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models*. Springer International Publishing.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.
- Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2), 281–301.
- Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1), 95–103.
- Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, 72(2), 264–290.
- Yen, W. M., Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.111–153). Praeger.

NOTE

1. Notice that this model does not include the specific factors on Items 1–4, unlike the preceding bifactor models. This model is a direct replication of the latent growth model in Chung et al. (2014).

Scaling, Equating, and Linking

Tim Moses

Buros Center for Testing

Few wish to assess others,
Even fewer wish to be assessed,
But everyone wants to see the scores
(Paul Holland, quoted in Dorans, 2018)

There are many ways that a testing program can assign and report scores that reflect test performance. Score reporting can be based on performance standards set by content experts; on relative standing with respect to one or more test-taker groups; on performance, strengths, and weaknesses across a set of tests contained in a test battery; or on performance for a particular age group or grade level in a scale constructed to track growth in age- or grade-appropriate tests of increasing difficulty. It would be difficult to overstate the importance of reported scores for a large-scale testing program. As a testing program's most visible and widely used products (Dorans, 2002; Kolen, 2006; Zenisky et al., this volume), it is essential to understand and document how reported scores are produced, maintained, and, at times, related to the scores reported by other testing programs. This chapter focuses on procedures for producing, maintaining, and linking testing programs' reported scores.

This chapter is an updated version of prior writings on scaling, equating, and linking. The attempt is to build on the seminal and foundational work of the chapters in previous volumes of *Educational Measurement* (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Kolen, 2006; Petersen et al., 1989). Extensive reviews of these chapters are, regrettably, not possible here, but are strongly encouraged. The major goal of the current chapter is to provide a "point-in-time" description of how issues and problems in the current testing field engage with, challenge, and build on prior chapters' discussions and frameworks for scaling, equating, and linking a testing program's reported scores.

- *Scaling* focuses on the theory and practice of establishing a testing program's reporting score scales.
- *Equating* focuses on the definition, requirements, history, and methodological practices of equating for maintaining a testing program's reporting score scales.
- *Linking* types are summarized with historical and recent examples for how a testing program might relate its reported scores to other scales.

In addition to the scaling, linking, and equating chapters in previous *Educational Measurement* volumes, the current chapter reflects two major influences. First, the testing program from which reported scores are produced represents a system of test production, administration, scoring, using, and interpreting test results, in multiple testing sites, and repeating in cycles over multiple points in time (Dorans, 2011; Holland, 1994, 2008). Understanding reported scores requires an understanding of the testing program's system, especially test specifications and development, administration conditions, and scoring procedures. The second influence pertains to the interpretations of reported scores. Testing standards emphasize that reported scores should be produced and maintained in ways that encourage

appropriate interpretations and discourage misinterpretations (American Educational Research Association [AERA] et al., 2014; Kolen, 2006; Petersen et al., 1989). The discussions in this chapter are intended to specify and elaborate on what this means and on ways to most effectively reflect “the overarching consideration . . . that users be given appropriate guidance about score interpretation and use” (Brennan, 2007, p. 175).

SCALING

The purpose of scaling is to establish the reporting scale(s) for the individual measures of a new or redesigned testing program. Scales are produced in the context of several activities of a testing program, such as the development of a test or measure from an established set of specifications for the content, construct, and individual items. Scaling is used in several testing contexts, including:

- admissions testing, where testing companies develop tests with relatively long scales for use in admissions decisions;
- K–12, where testing companies develop content-based tests associated with specific curricular standards adopted by policy makers to classify students at one or very few cut points for school and educator accountability, high school exit requirements, and instructional decisions;
- certification and licensure programs, where test content is determined with extensive input from practitioners and test takers are classified at cut points that indicate sufficient knowledge and skills to qualify for professional practice; and
- large-scale survey assessments developed by policy makers, educators, measurement and content experts, and other stakeholders for use in the estimation of trends in scale score distributions and classifications of populations and subpopulations over time.

Scaling activities also involve the administration of the developed test to test takers from a defined testing population under specific administration conditions, the scoring of test takers’ test items, and the conversion of test takers’ item scores into overall test scores or ability estimates. The data collection design for a test-taker sample from population p who take test form Y is shown in the first row of Table 11.1 (the terms and other designs in Table 11.1 are described throughout this chapter). The task in scaling is to develop a transformation that assigns numbers or ordered indicators to the test performance data.

The scaling process reflects the tests and their intended measurement, interpretations, and perspectives reviewed in the “Scaling Perspectives” section. Consider Y as a test form and a set of performance values test takers might receive, such that specific performance values are denoted $y = 0, 1, 2, \dots$. The y s can reflect sums of correct item scores, weighted summed scores, summed predicted probabilities of item-level models, or other indicators as described in the *Basic Unit of Scaling* section. Scale score transformations of

Table 11.1 Summary of Data Collection Designs for Scaling, Equating, and Linking

Description	Design Table					
	Population	Sample	Y			
Scaling/norming for a single test	P	1	$\sqrt$			
Single group	Population	Sample	X	Y		
	P	1	$\sqrt$	$\sqrt$		
Randomly equivalent groups	Population	Sample	X	Y		
	P	1	$\sqrt$			
	P	2		$\sqrt$		
Counterbalanced	Population	Sample	X_1	Y_1	X_2	Y_2
	P	1	$\sqrt$			$\sqrt$
	P	2		$\sqrt$	$\sqrt$	
NEAT	Population	Sample	X	A	Y	
	P	1	$\sqrt$	$\sqrt$		
	Q	2		$\sqrt$	$\sqrt$	
Common-item equating to a calibrated pool	Population	Sample	X_A	X_{New}		
	P	1	$\sqrt$	$\sqrt$		
Section pre-equating	Population	Sample	X_1	X_2	Y_1	Y_2
	P	1	$\sqrt$		$\sqrt$	$\sqrt$
	P	2		$\sqrt$	$\sqrt$	$\sqrt$
Item pre-equating	Population	Sample	X	New		
	P	1	$\sqrt$	$\sqrt$		

Note. NEAT = nonequivalent groups with anchor test. Depending on the context, X and Y can refer to alternate forms of a single test (equating) or to distinct tests (linking). For all but the NEAT design, P is assumed to be the target population for the test (T).

test performance are developed to produce reported scale scores that reflect increasing levels of achievement or proficiency, $sc(Y)$, that encourage specific interpretations. Two types of scales might be developed from the inputs, *primary* and *auxiliary*. Primary scales refer to reporting scales and test performance indicators that underlie the psychometric operations of a test (Kolen, 2006; Petersen et al., 1989), including the equating of scores from alternate test forms (see the *Equating* section) to the primary scale. Methods of establishing primary scale scores are described in the section *Scaling Methods for Primary Scale Scores*. Auxiliary scales are developed to facilitate interpretations and convey meaning in primary scales, and are described in the section *Auxiliary Scales*.

Scaling Perspectives

Early antecedents for educational and psychological testing and scaling include Horace Mann's efforts in 1845 to standardize oral exams of test-taker placement to control for the influence of different topics, examiners, and other efforts in psychophysics to apply scientific methods to study relationships between psychological sensations and physical stimuli (Briggs, 2022; Mislevy, 2018). Efforts to establish scales for nonphysical, psychological measures reflected these antecedents, as well as aspirations that measures and scales of nonphysical attributes might exhibit the measurement and scaling properties of physical attributes. The Binet–Simon Intelligence Test, released in France in 1905 by Alfred Binet and Theodore Simon, was an attempt to identify children in need of special services in ways that would be objective and free of teacher judgment and variations in the terminology, evidence, and reasoning used at that time (Binet & Simon, 1916; Cronbach, 1949). In the Binet–Simon test, test takers between ages 3 and 13 completed a series of 10 to 30 age-specific tests and were given a scale score designating an age and “intellectual level” corresponding to the test they were able to complete (Becker, 2003; Cronbach, 1949):

The fundamental idea of this method is the establishment of what we shall call a measuring scale of intelligence. This scale is composed of a series of tests of increasing difficulty, starting from the lowest intellectual level that can be observed, and ending with that of average normal intelligence. Each group in the series corresponds to a different mental level. (Binet & Simon, 1916, p. 40)

In another early example of scaling and measurement, Thorndike (1910) proposed a scale for handwriting quality by obtaining rankings of subsamples of handwriting samples of children in the fifth to eighth grades from competent judges and averaging these rankings to produce a scale for the entire sample. Thorndike (1910) described his scale of handwriting quality as one that ranged from “better than which no pupil is expected to produce, down to a quality so bad as to be intolerable, and probably almost never found, in school practice in the grammar grade” (p. 89). He also considered his scale of handwriting quality a way to assess educational outcomes similar to physical and scientific measures in other disciplines:

In general, the experience of constructing this scale gives great encouragement to the hope that for many educational facts, units and scales may be invented that shall enable us to think quantitatively in somewhat the same way that we can about facts of physics, chemistry or economics. (Thorndike, 1910, p. 150)

Scaling, measurement processes, and models were developed beyond the work of Binet and Simon (1916) and Thorndike (1910). For the Binet–Simon tests (i.e., items or questions) that were designed to be taken by children of specific ages, Thurstone (1925) developed a method to place the statistics of all Binet–Simon tests onto a single, continuous, normally distributed ability scale. This scale allowed Thurstone to evaluate and critique the *a priori* age groupings of the tests: “The questions are unduly bunched

at certain ranges and rather scarce at other ranges” (Thurstone, 1925, p. 448). Thurstone (1926, 1927) also provided a rationale and more fully developed theory of Thorndike’s (1910) scale in his law of comparative judgment (Coombs et al., 1970). This allowed for scales to be produced from any process of pairwise comparisons through modeling averages of pairwise differences and assuming they followed normal distributions. Guttman (1944) proposed a scaling approach that predicts an individual’s agreement/disagreement with an ordered set of unidimensional attitude statements in a deterministic model asserting that individuals who agree/disagree with a stronger/milder statement of attitude will also agree/disagree with all milder/stronger statements. Rasch (1960) proposed a scaling procedure based on a statistical logistic item response model of the probability of an individual’s response to an item based on one parameter for each item’s difficulty and one parameter for an individual’s ability (see the *IRT Ability Estimates* section; Wright, 1977).

Despite the methodological differences in previously described scaling approaches, comparative summaries have pointed out several similarities. The scaling approaches of Thorndike (1910) and Thurstone (1926, 1927) reflect an assumed “monotonicity property” in the nonphysical variables being scaled, implying consistency in judgments and subjective differences between stimuli (Coombs et al., 1970, p. 42). Thurstone’s, Guttman’s (1944), and Rasch’s (1960) approaches assume invariant comparisons in scales intentionally constructed to be unidimensional (Andrich, 1988). Engelhard (1984) compared the item calibration approaches of Thurstone (1925) and Rasch (1960) and another approach from Thorndike (1919), showing that these approaches produce similar empirical results and that they all attempt to eliminate the effects of samples and groups and thus achieve invariance. These scaling approaches can be described as giving increasingly greater emphasis to a particular model over data (Engelhard, 1984, p. 35). An implication of these approaches’ emphases on models is that uses and interpretations of the scales occur “only after a scale is developed that adequately fits the model” (Kolen & Brennan, 2014, p. 373).

Scaling has been addressed in fundamental definitions of what constitutes a measure and its scale. Early work was based on a classical perspective of discovering or estimating quantities and numerically representing them (Campbell, 1928; Hölder, 1901; Michell, 2008). From this perspective, attributes are established as quantitative and fundamental measures to the extent that they exhibit numerical properties that have a physical analogue, like length and concatenation (the adding or laying of objects end to end, Coombs et al., 1970; Hölder, 1901; Mislevy, 2018). An implication of these arguments is that measurement can be established only in the physical sciences. Psychophysical variables could not be established as measures because psychophysical variables would not exhibit concatenation operations or invariant relationships with established physical measures (Campbell, 1928; Ferguson et al., 1940).

Stevens (1946) provided a general typology of scales achieved through rules of assignment of numerals to a wide range of objects or events. By focusing on assignment rules rather than on requisite conditions for measurement (e.g., Campbell, 1928),

Stevens's (1946) scaling approach is more encompassing, alternatively described as paraphrasing (Stevens, 1946), circumventing (Mislevy, 2018), shifting (Briggs, 2022), or deflecting (Michell, 2008) the emphasis of earlier arguments on classical measurement and quantitative measures. For Stevens, scales are achieved in an operationalist (vs. classical) process of following (vs. discovering) rules for assigning numerals to objects according to their assumed level of measurement (vs. attributes of objects that must first be established as measurable; Briggs, 2022). The objects and measurements for which scales can be established include nominal (categorical labels), ordinal (ordered labels), interval (ordered labels that have equal intervals), and ratio (interval scales with a natural zero). In Stevens's arguments, the meaning and interpretation of scales depends on using admissible transformations for types of measurement (i.e., a monotonically increasing transformation for ordinal, linear transformations for interval) and on using permissible statistics for specific types of scales (e.g., medians for ordinal scales, means for interval scales). Stevens's scaling theory was further developed in subsequent work (Coombs et al., 1970; Krantz et al., 1971; Suppes & Zinnes, 1963).

Other approaches re-emphasized the classical focus on quantitative (interval or ratio) measurement through conditions other than concatenation (Coombs et al., 1970), including conjoint measurement (Luce & Tukey, 1964). One condition of conjoint measurement is double cancelation, which refers to a variable that consistently, additively, and noninteractively increases or decreases with two other variables (see the *Ability Estimates From Rasch and Other 1PL IRT Models* section).

Scaling approaches used in practice and described in measurement theory have been nearly separate pursuits in educational and psychological testing: "The axiomatic analysis of measurement models does not always provide feasible methods for constructing scales" (Coombs et al., 1970, p. 31). Binet and Simon (1916) described inconsistencies in the actual and aspirational measurement properties of their Binet–Simon test, simultaneously arguing for treating the scale of their test as equal to a quantitative measure even as they acknowledged that it reflected an ordinal scale of ordered, discrete classes:

This scale properly speaking does not permit the measure of intelligence, because intellectual qualities are not superposable, and therefore cannot be measured as linear surfaces are measured, but are on the contrary, a classification, a hierarchy among diverse intelligences; and for the necessities of practice this classification is equivalent to a measure. We shall therefore be able to know, after studying two individuals, if one rises above the other and to how many degrees, if one rises above the average level of other individuals considered as normal, or if he remains below. (pp. 40–41)

Another argument for treating ordinal measures as though they reflected interval scales in scaling practice appealed to pragmatics and usefulness over formal measurement properties:

Although, formally speaking, interval measurement can always be obtained by specification, such specification is theoretically meaningful only if it is implied

by the theory and model relevant to the measurement procedure. At various times in this book however, we shall treat a measurement as having interval scale properties, although it is clear that the measurement procedure and the theory underlying it yield only a nominal or, at best, an ordinal scale . . . from a pragmatic point of view, the only meaningful evaluation of this procedure is one based on an evaluation of the usefulness of the resulting scale. (Lord & Novick, 1968, pp. 21–22)

Some objections have been expressed about Stevens's (1946) arguments concerning levels of measurement and permissible statistics, such as Lord's (1953) argument that statistical operations seemingly reserved for interval scales could be appropriate even for nominal scales like numbers on football players' jerseys. Connections of measurement theory and scaling practice have been considered in terms of the statistical and probabilistic Rasch models and conjoint measurement (Embretson & Reise, 2000; Karabatsos, 2001; Perline et al., 1979). However, noted limitations in the fit of the Rasch model have raised questions about the inconsistencies of pursuits to establish scales that reflect measurement theory versus those that fit actual test data (Briggs et al., this volume; Embretson, 2006; Embretson & Reise, 2000; Thissen & Orlando, 2001; Wright, 1994).

The perspectives taken in describing the practice of establishing reporting scales for educational tests in the first four volumes of *Educational Measurement* make specific references to, and departures from, previous discussions of scaling and measurement (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989). Most of the *Educational Measurement* chapters present a basic definition of scaling similar to Stevens's (1946), one of assigning numerals or numbers to a test taker's test performance. Interpretation and meaning are less about requirements for measurement, statistics, or scaling transformations and more about the scale properties that might be desirable for other criteria and interpretations (described in the *Scaling Methods for Primary Scale Scores* and *Auxiliary Scales* sections). All of the *Educational Measurement* volumes discussed scale score interpretations in terms of scale score distributions on relevant testing populations (norms). Discussions of interval scales and other measurement issues either appear in other nonscaling chapters (e.g., Lorge, 1951) or are described in terms of their limited usefulness for scaling and psychological measures due to the absence of operational definitions that are agreed on by experts (Angoff, 1971) and the lack of a complete theory and definition for the psychological and educational constructs in need of scales (Kolen, 2006). Finally, perspectives on scaling for educational tests emphasize a starting point that differs from the one implied in measurement definitions and models. Petersen et al. (1989), Kolen (2006), and Brennan (personal communication, September 17, 2020) have distinguished educational measurement from other types of measurement because the starting point in educational measurement, that is, test content and specifications, comes from an external entity (e.g., one of the testing contexts summarized in the beginning of *Scaling*). They argue, consistent with Lindquist (1953), that

the objective is handed down . . . by those agents of society who are responsible for decisions concerning educational objectives, and what the test constructor must do is to attempt to incorporate that definition as clearly and as exactly as possible. (p. 35)

For educational measurement, content, test development, and measurement models influence the development of a test's reporting scale in choices for the basic unit used in scaling (see the *Basic Unit of Scaling* section), transformations of test performance measures to a reporting scale with desired properties (see the *Scaling Methods for Primary Scale Scores* section), and maintenance of the scale in subsequently developed forms through equating (see the *Equating* section).

Basic Unit of Scaling

The *Scaling Perspectives* section described fairly different views on scaling. These differences are apparent in the options for defining the basic unit of test performance used to establish a testing program's score scale. This section covers choices and implications for representing test performance, including different types of observed test scores that summarize observed test performance, and the latent abilities estimated using some item response theory (IRT) models. For other summaries of these options, including descriptions of approaches for scoring different types of items, see Kolen (2006) and Dorans (2018).

Observed Test Scores

Test performance indicators can be established as different types of observed scores. For test form Y that contains $i = 1$ to I items that are not distinguished with respect to content area or item format, the most common observed score is the sum of the $i = 1$ to I item scores, V_i

$$Y = \sum_i w_i V_i \quad (1)$$

where $w_i = 1$ for all I items. With this option, a test taker's score reflects equally weighted contributions from each item in the test. Other options involve the summing of differentially weighted item scores, where the item score weights might be chosen to maximize some measure of test reliability or validity (Gulliksen, 1950; Lord & Novick, 1968). Although these reliability and validity measures for maximization are defined by Gulliksen (1950) and Lord and Novick (1968), many issues concerning reliability, true score variance, and error variance for scaling procedures are, at best, imperfectly understood. Other texts should be consulted for comprehensive discussions of reliability (Brennan, 2001; Haertel, 2006; Lee & Harris, this volume). Different options suggested in Equation 1 can be used to summarize test takers' test performance on a specific set of test items.

Test form Y may also be a "composite" or "mixed format" test containing items with different content areas and/or item formats. When Y includes I multiple-choice (MC)

items and J constructed-response (CR) items, an approach to weighting the item type scores and obtaining observed composite scores is

$$Y = w_{MC} \sum_i V_{i,MC} + w_{CR} \sum_j V_{j,CR}. \quad (2)$$

One option is to nominally weight the scores from the MC and CR items so that the weighted scores reflect an intended number and percentage of composite score points. The Advanced Placement exams provide several simple and complex examples of MC item score weights and CR item score weights being set such that the weighted and summed scores contribute desired numbers of points to the composite score (College Board, n.d.; Moses et al., 2006). Another option is to determine effective weights such that the weighted scores reflect a desired proportion of the composite's observed score variance or true score variance. A third option is to select weights for the scores summed from different item formats such that some measure of composite score reliability is maximized.

The options for summing item scores produce measures of test performance with different implications. Sums of item scores that are equally weighted or weighted to reflect intended numbers of composite score points are the simplest methods of reporting test performance, associated with scoring rules that can be fairly easily communicated to test takers (i.e., "To do as well as possible, answer every item"). Simple summed scores can directly reflect expert judgments about the test as indicated by the numbers and weights of different items described in test specifications. Other options weight items based on statistics such as variances, maximized reliabilities, or validities and have interpretations based on those criteria. In addition, the weights reflect population and sample characteristics of those corresponding statistics. The scoring rules associated with these options for weighted observed test scores are more complex and may be more difficult to communicate to test takers depending on when these population-dependent weights are derived (i.e., test takers may approach the test in a way that is suboptimal or inconsistent with item weights that may be derived before or after the test is administered). Most of these observed score options should be understood to reflect an unspeeeded test where test takers have sufficient time to answer every item. Inadequate testing time can elicit rushed or random responding that could result in unintended measurement properties for those speeded items and the overall test. Equally or nominally weighted item scores may be suboptimal with respect to test score reliability, especially if these weights result in higher contributions from the scores of less reliable item types (i.e., higher contributions of CR scores relative to MC scores).

Observed test scores are direct indications of actual test performance and indirect reflections of unobservable latent abilities. As such, potential uses of observed test scores as estimates of latent abilities require supplemental measures of the reliability or generalizability of the scores to other possible item samples, test administrations, and admissible measurement conditions. When interpreted in terms of measurement and scale properties, test scores are often described as ordinal scales for some latent ability they are assumed to measure (Embretson & Reise, 2000). Alternatively, observed test scores could be considered interval scales from a pragmatic perspective

(Lord & Novick, 1968), or from a perspective where scores from a particular test form are regarded as indices of test performance rather than estimated abilities, such that a test form's scores exhibit equal increases with additional correct responses to test items (Mislevy, 2018, p. 316).

IRT Ability Estimates

Scales might be established on estimated latent abilities rather than on the test performance observed for a set of test items. In theory, the item response data from a test form might be modeled in ways that produce estimates of abilities that are independent of the test form. In practice, this involves fitting IRT models to data from a test and its fixed, specific set of items, such that the item characteristics require scaling procedures to account for their sample effects (described by Kolen & Brennan, 2014). The result is that test performance measures can be produced that may be regarded as estimates of latent ability, but that are essentially computationally complex summaries of observed item performance, in need of their own reliability estimates (Lee & Harris, this volume).

Assume test form Y contains dichotomously scored items with correct responses scored 1 and incorrect responses scored 0. Assume further that test takers' responses to the I items are conditionally independent given a latent and unidimensional ability, θ , and that the probability of a correct response follows a logistic model with up to three parameters, such that the probabilities monotonically increase with θ . The item response probabilities from the one-, two-, or three-parameter logistic models (1PL, 2PL, 3PL) can be expressed as

$$Pr(V_i = 1 | \theta, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + \exp[-Da_i(\theta - b_i)]} \quad (3)$$

where a_i , b_i , and c_i are discrimination, difficulty, and guessing parameters for item i and D is a scaling constant that is sometimes 1 and other times set to 1.702 so that Equation 3's logistic function approximates a normal ogive (Haley, 1952). This section summarizes some approaches for estimating test takers' abilities, θ , based on test takers' scored responses to the items on test form Y , $v_1, v_2, \dots, v_I$ and on parameter estimates for all I items on the test form that are treated as population values (not estimated). IRT ability estimates reflect several choices for fitting IRT models to test data, including calibration decisions, estimation software, and approaches for polytomously scored items and mixed format tests (for additional discussions of IRT models, see Cai et al., this volume).

Three commonly used IRT ability estimates are the maximum likelihood estimate (MLE), the Bayesian expected a posteriori (EAP), and the test characteristic curve (TCC) estimate (for others, see Thissen & Orlando, 2001).

The MLE is obtained by solving for θ using an iterative procedure to maximize,

$$L(V_1 = v_1, V_2 = v_2, \dots, V_I = v_I | \theta) \\ = \prod_i Pr(V_i = 1 | \theta, a_i, b_i, c_i)^{v_i} [1 - Pr(V_i = 1 | \theta, a_i, b_i, c_i)]^{(1-v_i)}. \quad (4)$$

The MLE can be described as the ability estimate with maximum information (i.e., precision) and inverse sampling variance. The MLE corresponds to an optimally weighted observed score, $\sum_i w_i V_i$, for a particular IRT model (Lord, 1980; Yen & Fitzpatrick, 2006).

The Bayesian EAP estimate,

$$EAP = \frac{\int_{\theta} \theta L(V_1 = v_1, V_2 = v_2, \dots, V_I = v_I | \theta) Pr(\theta) d\theta}{\int_{\theta} L(V_1 = v_1, V_2 = v_2, \dots, V_I = v_I | \theta) Pr(\theta) d\theta}, \quad (5)$$

is the mean of a distribution of θ obtained as the product of the likelihood and an assumed prior distribution for test-taker ability, $Pr(\theta)$. In practice $Pr(\theta)$ is usually represented as a discrete approximation of the standard normal distribution, though it can also be obscure.

The TCC estimate is based on relating the IRT-expected summed scores given θ , $\tau_Y(\theta)$, to the observed summed score (Equation 1). Ignoring measurement error in the scores of test form Y and making other adjustments to account for true scores that are undefined in the IRT model, Equation 1 and a summed version of Equation 3 are set equal to each other and θ is estimated to preserve this equality:

$$Y = \sum_i w_i V_i = \sum_i w_i Pr(V_i = 1 | \theta, a_i, b_i, c_i) = \tau_Y(\theta) \quad (6)$$

Although the solution to θ in Equation 6 ensures that the IRT-expected summed score matches the observed score, this application of true score relationships to observed scores lacks justification (Kolen & Brennan, 2014, p. 201).

ABILITY ESTIMATES FROM RASCH AND OTHER 1PL IRT MODELS For the Rasch (1960) (all $a_i s = 1$, all $c_i s = 0$, and $D = 1$) and other 1PL models (all $a_i s = \text{constant}$, all $c_i s = 0$, and $D = 1.702$), ability estimates have a one-to-one relationship with the test score summed from equally weighted items, all $w_i s = Da_i$, and the TCC and MLE estimates are equal (Thissen & Orlando, 2001). When a uniformly distributed prior is used, $Pr(\theta)$, EAP ability estimates are also equal to those of MLE and TCC, except that the highest and lowest summed scores have estimates from EAP but are undefined with MLE and TCC. These results reflect a property of Rasch and 1PL models that the summed test score with equal $w_i s$ is a sufficient statistic that contains all information needed for estimating θ (Lord, 1980). Response probabilities from Rasch and 1PL models are additive and consistently ordered for item difficulties (or test-taker abilities), regardless of abilities (or items), always increasing as item easiness and test-taker ability increase. This means that the Rasch and 1PL models exhibit double cancelation and could be said to have a probabilistic relationship to conjoint measurement (Briggs et al., this volume; Embretson & Reise, 2000; Mislevy, 2018; Perline et al., 1979). Although this relationship has been the basis of attempts to establish quantitative measures with interval properties in test data, these attempts are challenging because of model–data fit limitations (see the *Scaling Perspectives* section). The relationship also reflects inconsistencies in

observed item responses, the probabilistic Rasch (1960) model of latent and estimated item parameters and abilities, and the deterministic framework of observed variables in conjoint measurement (Kyngdon, 2008; Mislevy, 2018; Perline et al., 1979).

As stated in the *Scaling Perspectives* section, Rasch models are more restrictive and less likely to fit observed data than other IRT models. These model–fit limitations can prompt calls to impose statistical consequences to the data or edit or remove nonfitting responses (Wright, 1977). Examples include “reinterpreting or modifying the frame of reference” (Andrich, 1988, p. 62) or reducing the intended scale:

The broader the domain of interest, the more difficult it will be to make targeted and testable hypotheses. This would suggest that vertical scales could only be plausibly supported for more narrowly defined latent variables. (Briggs, 2013, pp. 219–220)

Removal of test-taker data could potentially limit intended scale interpretations, such as by narrowing the intended scaling population (i.e., changing P in Table 11.1). Removal of nonfitting items could result in narrowing the construct defined in test specifications.

ABILITY ESTIMATES FROM 2PL AND 3PL IRT MODELS The 2PL (unique a_i s and c_i s = 0) and 3PL (unique a_i s and c_i s > 0) IRT models can be used to produce more complex ability estimates with unique contributions from individual items. The MLE and EAP estimates from 2PL and 3PL models reflect patterns of correct and incorrect responses to individual items, such that they differ from and convey more information than the TCC estimates. The MLE for the 2PL IRT model reflects item pattern scores and an optimally weighted test score where item scores are weighted by their discrimination, $w_i = Da_i$, and the resulting weighted test score is the sufficient statistic for θ . A TCC procedure usually based on equally weighted item scores differs from the MLE, but can produce the MLE estimate when modified to sum optimally weighted item scores (Lord, 1980). For the 3PL model, there is no sufficient statistic for θ , and the MLE estimate reflects a test score summed using optimal weights defined

$$\text{as } w_i = \frac{Pr(V_i = 1 | \theta, a_i, b_i, c_i) - c_i}{1 - c_i} \frac{Da_i}{Pr(V_i = 1 | \theta, a_i, b_i, c_i)}.$$

The MLE, EAP, and TCC ability estimates based on 2PL and 3PL IRT models have unique properties (S. Kim & Moses, 2025; Kolen, 2006; Kolen et al., 2011; Thissen & Orlando, 2001; Yen & Fitzpatrick, 2006). MLEs (Equation 4) are asymptotically unbiased estimates of θ , though for tests of realistic length they have conditional biases that are positively correlated with θ (Lord, 1983) and an overall variance that is larger than that of θ . MLE estimates are unavailable for some response patterns, such as those where the responses to the test items are all incorrect or all correct. EAP estimates from Equation 5 have conditional biases that are negatively correlated with θ (Lord, 1986), and they are less variable than the MLEs. The EAP estimates reflect shrinkage to the prior distribution of θ , meaning that θ values are overestimated for test takers below the mean in the population and underestimated for test takers above the mean. θ estimates based on EAP are available for all item response

patterns. Ability estimates based on EAP and MLE item pattern scores have smaller standard errors than those using the usual TCC approach, though their similarity with TCC estimates increases for longer tests with items that have smaller guessing effects (Yen & Fitzpatrick, 2006).

Ability estimates for 2PL and 3PL IRT models have implications for model–data fit, theoretical measurement properties, and score interpretations. These IRT models fit observed test data more closely than Rasch and 1PL models, but do so with added complexity. From a theoretical measurement perspective, interval properties in ability estimates from 2PL and 3PL IRT models are more difficult to defend than for Rasch models. Item characteristic curves for given items can cross for higher and lower test-taker abilities, and this nonadditivity means that 2PL and 3PL IRT models are not guaranteed to meet the double cancellation condition or have a probabilistic relationship to conjoint measurement. Ability estimates are nevertheless sometimes described as interval scales because 2PL and 3PL IRT models produce probabilities that are invariant under different linear transformations of their parameters for a specified functional form and population distribution (contrasting points are made about this description in Briggs et al., this volume, endnote 10, and Mislevy, 2018, pp. 316–317). Empirical research on interval scales from these IRT models has been encouraged (Michell, 2008), and statistical tests may be possible for the 2PL model (Kygndon, 2011).

From an interpretational perspective, MLE and EAP ability estimates from 2PL and 3PL models are more complex and more difficult to understand than those based on simple sums of equally weighted item scores (i.e., the usual TCC). The weighting of individual items in the scoring is likely to be less closely aligned to the intended test content as determined by test developers and described in test specifications (Kolen, 2006). The complexity of ability estimates based on pattern scoring implies more complicated scoring rules and increased difficulty in communicating, understanding, and explaining scoring and advising on how test takers might maximize their performance. For MLEs based on the 2PL model, items with higher discriminations make greater contributions to test takers' ability estimates, which creates fairness issues when these differentially important items are not communicated to test takers (Dorans, 2012). For the 3PL model, MLEs reflect differentially weighted items for different test takers (i.e., for less able test takers, correct item responses tend to count less and are more likely attributed to guessing; Lord, 1980, p. 75). This characteristic was described by Mislevy (2012) as “unsettling” and tough to explain (p. 39). The fairness of pattern scoring based on the 3PL model is questionable and has been described as a challenge for establishing comparability in adaptive tests (Phillips, 2016, p. 258). A complication with EAP estimates is that they reflect not only test performance in item pattern scores, but also a test-taker ability distribution (Kolen, 2006).

ABILITY ESTIMATES FROM MORE COMPLEX MODELS Scoring approaches have been developed and considered that are even more complex than those that have been previously described. These approaches are developed based on a range of motivations. One

is to obtain scores on noncognitive traits assessed with faking resistant forced-choice item formats using item parameter estimates obtained in previous administrations of those items in formats that are not faking resistant (Drasgow et al., 2012; Stark et al., 2005). Other efforts score tests composed of family-generated clone items using IRT methods like Equations 4–6 based on parameters from the item families rather than parameter estimates for the items test takers actually take (cautions provided in Drasgow et al., 2006; Harris, 2023; Luecht & Burke, 2020; van der Linden & Glas, 2010). Other efforts are being considered for measuring response processes like test-taking strategies, thought processes, and behaviors like reading, interpretation, and strategy formulation with response times, computer screen gazes, verbalizations, and other process information (Ercikan & Pellegrino, 2017). Deep-learning neural network models might be considered for automated scoring versions of subjective human scoring of test takers' writing (Zesch et al., 2023).

For complex scoring approaches, test takers' test performance measures might reflect not only their performance on the test and items they take, but also other unspecified aspects of the scoring procedures. Scoring based on item families can produce test-taker scores that reflect errors in the IRT parameter estimates of items they may not actually take. Complex scoring algorithms can be difficult to interpret and explain (Lottridge et al., 2023; Zesch et al., 2023), and the resulting test performance measures may reflect reduced accuracy and fairness, or unrepresentativeness in the data used to train the models (Broussard, 2020; Hussein et al., 2019; W. Lee & Harris, this volume). Performance measures that reflect test-taking processes can warrant validation efforts and attention to the processes different test takers use to optimize their performance (Kane & Mislevy, 2017; Wise, 2017). These approaches to representing test performance are mentioned here to illustrate recent trends for increased complexity and, likely, increased difficulty in explaining, justifying, and defending scales for the resulting performance measures.

Scaling Methods for Primary Scale Scores

The approaches to representing test performance described in the *Basic Unit of Scaling* section are usually considered inadequate for use as a testing program's reporting scale (Angoff, 1971; Kolen, 2006). Observed test scores represent test performance in ways that are specific to a test form, its items, and its measurement characteristics (difficulty, reliability, etc.). Under strong assumptions, IRT ability estimates may be regarded as theoretically independent of a test form and its items, but even if these assumptions are met in practice, the ability estimates have their own interpretational difficulties because of their similarity to standard normal variables, with means near zero, and small numbers with decimals that can be negative or positive. Consistent with the scaling discussions from earlier *Educational Measurement* volumes (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989), scale scores established from untransformed observed scores, $sc(Y) = Y$, or untransformed estimated IRT thetas, $sc(\hat{\theta}) = \hat{\theta}$, are not recommended.

The task in scaling is to develop a numeric transformation of a measure of test performance that assigns numbers or ordered indicators such that the scale scores,

$sc(Y)$, reflect monotonically increasing levels of achievement. For large-scale testing programs, this task of establishing scale transformations is intended to be applied not only to the scores of test form Y , but also to other alternate versions of Y developed, administered, and equated (see the *Equating* section) to Y . Another goal of scaling is that the scale scores be established to facilitate score interpretations and discourage misinterpretations. In this section, different scale score transformations are reviewed that focus on the structure, normative, shape, measurement, and content aspects of reporting scales. In each subsection, methods and intended interpretations are described.

Reporting Scale Basics

Some aspects to be determined for a testing program's reporting scale are structural, such as the scale score range and number of possible points. Recommendations are usually to establish the range of the scale scores in such a way that the scores cannot be easily interpreted as observed test score performance (Dorans, 2002; Kolen, 2006; Livingston, 2004). For example, scale score ranges such as 100–200 (Praxis) and 200–800 (SAT) are wide enough to impede unwarranted guesses about the relationship between the number of items and scale score points. The GRE scale score range (130–170) and the ACT range (1–36) are also not likely to be interpreted as simple transformations of raw scores.

Another aspect to consider in scale scores is the interval, or the number of possible scale score points reflected in the range of the scale scores. Most recommendations are to establish the scale score interval such that it represents the available information or precision of the test (Dorans, 2002; Flanagan, 1951; Kolen, 2006; Livingston, 2004). When intervals are too fine, the resulting scores could lead users to overinterpret score differences, such as on pre-1970 SAT scales of 200–800 in integer units (Livingston, 2004). When scale score intervals are more coarse than test performance (e.g., stanines), then the scale scores can result in lost information (Flanagan, 1951; Kolen, 2006) or test takers with very different test performance receiving the same scale score (Dorans et al., 2010).

Scale score intervals might be established based on simple recommendations, such as ensuring that the possible scale score points do not exceed the number of observed score points on the test (Dorans, 2002) or establishing desired confidence intervals for true scores or scale scores (Kolen, 2006). Kolen (2006) summarized two proposals for establishing scale score points that reflect desired confidence intervals for true scores for the Iowa Tests of Educational Development (ITED, 1958) and more generally (Kelley, as cited in Kolen, 2006, p. 165). These procedures are based on assuming that measurement error is normally distributed and constant across the scale, that the reliability of test performance is known or well estimated in population P , and that the scale score transformation is linear. With these assumptions, a scale score transformation that reflects true score confidence intervals could be established based on

$$6\sigma_{sc(Y),P} = 6 \frac{h}{z_Y \sqrt{1 - rel_{Y,P}}}, \quad (7)$$

where h is the width of the desired scale score interval, z_γ is the standard normal z -score associated with the desired γ 100% confidence interval, $rel_{Y,P}$ is the reliability of Y in population P , $\sigma_{sc(Y),P}$ is the standard deviation of $sc(Y)$ in P , and $6\sigma_{sc(Y),P}$, when rounded, provides a recommended number of distinct scale score points. For the ITED implementation in Equation 7, the scale score units are set such that adding or subtracting one scale score point establishes a 50% confidence interval for true scores (i.e., $h = 1$ and $z_\gamma \approx .6745$ for $\gamma = 50\%$). For Kelley's implementation of Equation 7, the scale score units are set such that adding or subtracting three scale score points establishes an approximate 68% confidence interval ($h = 3$ and $z_\gamma \approx 1$ for $\gamma = 68\%$). From Kolen's (2006) summary of these proposals, Kelley's rule generally leads to about twice as many scale score points as the ITED rule. For both approaches, the number of recommended scale score points decreases as reliability decreases. Equation 7 can be used with any measure of test performance for which a reliability estimate is available, including an IRT θ . Once used, a linear scale score transformation is found that produces scale scores with a range of units that is consistent with Equation 7.

Another structural aspect of scale scores is the level of truncation, where recommendations are for the range of minimum and maximum reported scale scores to be narrower than the actual scale score range (i.e., the working range, Dorans, 2002). Truncation of the maximum scale scores has been recommended because it avoids interpretational difficulties such as conveying perfect test performance on test forms that differ in difficulty and have different untruncated scale scores (Livingston, 2004). Truncation also helps make the resulting scale scores more resistant to shifts in score distributions due to changes in the population or in the difficulties of the test forms (Dorans, 2002). Establishing a minimum reported scale score that is higher than what is suggested by the scale score transformation can be useful for avoiding meaningless distinctions at test performance levels lower than theoretical guessing levels where measurement is less precise (Livingston, 2004).

Finally, consider that a reporting scale can be established to specify the scale score that should correspond to one or more specific observed scores of Y . The situation of interest is one for which a scale score transformation is established such that two observed scores of test Y , y_1 and y_2 , will have prespecified scale scores, $sc(y_1)$ and $sc(y_2)$:

$$sc(y) = sc(y_1) + (y - y_1) \frac{sc(y_2) - sc(y_1)}{y_2 - y_1} \quad (8)$$

This transformation has been described as a basis for a scale score range of 100–200 on Praxis (Livingston, 2004).

Normative

When establishing scale scores, a common desire is for the scale scores to represent a test-taker group of interest in terms of one or more statistics (i.e., statistics of a reference

group or sample of population P taking test Y ; Table 11.1). Scale score conversions might involve transformations of the observed test performance on Y to a specified scale score value. One commonly used linear transformation converts the mean and standard deviation of test performance in population P , $\mu_{Y,P}$ and $\sigma_{Y,P}$, to desired scale score values, $\mu_{sc(Y),P}$ and $\sigma_{sc(Y),P}$:

$$sc(y) = \mu_{sc(Y),P} + (y - \mu_{Y,P}) \frac{\sigma_{sc(Y),P}}{\sigma_{Y,P}}. \quad (9)$$

Equation 9 can be used to establish raw-to-scale score transformations that satisfy several scaling criteria. For example, to satisfy a recommendation that the scale score mean be set at the center of the possible scale score range (Dorans, 2002), Equation 9 might be used to set the mean test performance for a test-taker group to be $sc(y) = 500$ for a 200–800 range (SAT; Dorans, 2002) or 18 for a 1–36 range (ACT; Brennan, 1989). Also, Equation 9 might be used to produce a set of viable raw-to-scale score transformations that vary in how they satisfy different criteria for numbers of distinct scale score points (Equation 7) and how far the actual scale scores extend below and above a truncated range of scale scores, as well as to establish the means and standard deviations for nonlinear raw-to-scale score transformations (reviewed next).

Another example of a linear scale score transformation that reflects the normative information for a group of test takers is to establish a desired scale score standard deviation and a conversion of one particular score, y_1 , to $sc(y_1)$,

$$sc(y) = sc(y_1) + (y - y_1) \frac{\sigma_{sc(Y),P}}{\sigma_{Y,P}}. \quad (10)$$

where $\sigma_{sc(Y),P}$ is the intended standard deviation of the $sc(y)$ s in population P . Linear raw-to-scale transformations such as Equations 9 and 10 establish scale scores as linearly increasing indications of test performance, retaining the shape (skewness and kurtosis) of the raw scores in the scale scores.

Nonlinear Transformations

Nonlinear transformations can be used to establish scale scores that encourage interpretations based on scale score basics (see the *Reporting Scale Basics* section), distribution shape, or measurement precision. Rounding scale scores to integers or other units can establish scale scores that convey an intended range of possible scale score points. Truncating the lowest and highest possible scale score values to intended values also helps to consistently maintain the structural aspects of scale scores. Nonlinear raw-to-scale score transformations might also be developed to satisfy other criteria. These might be subjected to an additional linear transformation like Equation 9 to reflect aspects such as the intended range and number of possible scale score points.

One nonlinear transformation involves establishing scale scores that are approximately normally distributed. The resulting scale scores can then be described as symmetric and reflective of a shape that is familiar to test users (Dorans, 2002; Petersen et al., 1989). Normally distributed scale scores can be established by first obtaining the

percentile ranks or “percent at and below” for a measure of performance on test Y (Kolen & Brennan, 2014). For this discussion, consider the measure of test performance to be the raw scores obtained as the sum of I items, y_k , $k = 1$ to I , where the actual scores range from 0 to y_I . A continuized version of the scores is assumed where each individual score, y_k , ranges from $y_k - .5$ to $y_k + .5$ and the entire set of continuized scores ranges from $-.5$ to $y_I + .5$. The percentile ranks for these continuized scores can be computed as

$$\begin{aligned} G_P(y_k) &= 100 \left\{ \sum_{y_j < y^*} Pr_P(y_j) + \right. \\ &\quad \left. \left[y_k - (y^* - .5) \right] Pr_P(y^*) \right\}, \quad -.5 \leq y_k < y_I + .5 \\ &= 0, \quad y_k < -.5 \\ &= 100, \quad y_k \geq y_I + .5, \end{aligned} \quad (11)$$

where $Pr_P(y_k)$ is the relative frequency at score y_k for population P and where y^* is the closest integer to y_k such that $y^* - .5 \leq y_k < y^* + .5$. Values of the standard normal distribution are found, z_k , such that their cumulative distribution values, $\Phi(z_k)$, equal a function of the percentile ranks of the y_k s,

$$G_P(y_k) / 100 = \Phi(z_k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_k} e^{-w^2/2} dw. \quad (12)$$

In Equation 12, w is a variable that assumes values ranging from $-\infty$ to z_k in the integration denoted by $\int_{-\infty}^{z_k} \dots dw$. The resulting z_k s can be linearly transformed to have an intended mean and standard deviation (Equation 9) and rounded and truncated to have the intended range of possible scale scores. The normal transformation was used to establish recentered SAT scales in 1995 based on the 1990 reference group of graduating seniors (Dorans, 2002). Other examples provided by Kolen (2006) and Kolen and Brennan (2014) illustrate that the normal transformation is approximate and that its closeness to the normal distribution depends on the distribution of the raw scores.

Another motivation for nonlinear transformations is producing scale scores that provide a stabilized version of conditional measurement error. That is, testing standards call for reporting conditional standard errors of measurement (CSEMs) for several scores if CSEMs differ across the score range (AERA et al., 2014). For most psychometric models, the CSEMs of scores are small for the highest and lowest unscaled true scores and large for true scores in the middle of the range (Lord & Novick, 1968). Kolen (1988) proposed an arcsine transformation developed by Freeman and Tukey (1950) for use as a scale score transformation of number-correct true scores. Assuming that the errors of number-correct true scores follow a binomial, compound binomial, or IRT model (Kolen, 2006), scale scores having approximately equal CSEMs across their range can be produced as

$$g(y) = .5 \left\{ \sin^{-1} \left[\frac{y}{I+1} \right] + \sin^{-1} \left[\frac{y+1}{I+1} \right] \right\}. \quad (13)$$

The resulting scale scores can be reported and described as reflecting a single standard error of measurement (SEM) value across the range of true scores (Kolen et al., 1992). They can also be further transformed to have an intended mean and standard deviation (Equation 9) and rounded and truncated to have an intended range of possible scale scores. Arcsine transformations have been used to establish scales for the ACT (Brennan, 1989) and the 2016 redesigned SAT (Y. K. Kim et al., 2016).

Nonlinear transformations to achieve normal distributions or to stabilize CSEMs can have interpretational difficulties. Both types of transformations can result in extreme stretching of the highest and lowest scale scores, exaggerating scale score results and creating gaps (missing scale scores) for the highest and lowest levels of test performance. These methods can also produce scale scores with inconsistent properties, such that arcsine transformations can result in more skewness (less symmetry) in the scale score distributions. A conceptual issue with CSEM stabilization is that the scales are established for true scores and applied to observed scores.

To produce a scale score transformation that could achieve symmetry with less extreme conversions of the highest scale scores, Moses and Golub-Smith (2011) proposed a cubic transformation of raw scores,

$$sc(y) = \delta_0 + \delta_1 y^1 + \delta_2 y^2 + \delta_3 y^3, \quad (14)$$

where the δ s are derived to achieve a desired mean, standard deviation, skewness, and kurtosis in the scale scores (and also to satisfy a constraint that the scale scores are monotonically increasing). This procedure can be used to approximate the normal distribution, but to varying degrees that produce less extreme transformations and fewer gaps at the highest and lowest scale score regions. This procedure was used to set the scales of the revised GRE (Golub-Smith & Moses, 2014). Another version of the cubic transformation was proposed by Moses and Kim (2017) to produce scale scores that would stabilize an inputted set of CSEMs. This procedure was used to set one of the scales of the 2016 redesigned SAT (Y. K. Kim et al., 2016). Moses and Kim (2017) also showed how cubic transformations could be established that satisfy a set of scaling criteria to varying degrees, such as by making scale score CSEMs more similar while also targeting scale score symmetry. Finally, because CSEMs are inputs in developing the scale score transformation, the cubic transformation can also target stabilizations for a limited number of scores involving consequential decisions (i.e., cut scores; Lord, 1985) or for different measures of test performance, such as CSEMs at IRT θ s or at true scores and test performance measures for mixed format tests. Another situation where the flexibility of the cubic transformation might be utilized is stabilizing true score intervals for a specified range of observed scores (Lord & Novick, 1968), an application that would likely produce different scaling results than those produced to stabilize CSEMs given true scores. Finally, the cubic transformation could be extended to fit even higher moments in a targeted distribution (Headrick, 2002).

Interval Scales

Some discussions of psychological measures (Michell, 2008) and vertical scales (Briggs, 2013) have taken issue with claims that psychological and educational measures can be interpreted as quantitative and as having interval properties. A recommendation from these perspectives is that assumptions for the treatment or interpretation of variables as quantitative should be explicitly acknowledged or, better yet, should prompt empirical evaluations such as empirical tests of whether the variables meet conjoint measurement conditions (see the *Scaling Perspectives* and *Basic Unit of Scaling* sections, Michell, 2008). As suggested in the *Basic Unit of Scaling* section, most available measures of test performance are difficult to establish as quantitative scales for the above notion of “quantitative.” Some apparent consequences have been described in the context of vertical scaling (Briggs, 2013), where the intent is to establish a scale where growth across grades can be measured and interpreted with respect to interval interpretations (see the *Vertical Scaling* section). Claims about vertical scales exhibiting interval properties might be made by test publishers, claims that have been discussed as inconsistent with growth rates observed to vary across lower versus higher grades or across lower versus higher parts of the scale score distribution (Briggs, 2013; Hoover, 1984). Empirical evaluations for the existence of intervals in vertical scaling results have been suggested by Briggs and Domingue (2013).

The starting point for establishing a scale with interval properties is a measure of test performance with interval properties. From the *Basic Unit of Scaling* section there are at least three possibilities:

- operationalist or pragmatic perspectives of the observed test scores as interval-based indices of performance for a specific test form
- Rasch ability estimates from test data shown to fit a Rasch model, specifically with empirically demonstrated double cancelation and conjoint measurement conditions
- ability estimates from 2PL or 3PL models, based on arguments about the specification and invariance of the probabilities from these models under different linear transformations

Following Stevens (1946), linear scale transformations (Equations 8, 9, and 10) of these test performance measures would retain the established, or argued-for, interval properties, whereas nonlinear transformations would not. Note that the linear transformation would apply to one of the three possibilities but not to all three, because the three possibilities cannot coexist because of inconsistencies in the assumed model (Rasch vs. 2PL or 3PL) and nonlinear relationships in observed test scores and IRT-based abilities.

Scales for Exams With Cut Scores

For K–12 and certification and licensure tests, the interest is in reporting performance for a limited number of performance categories rather than on the entire range of a score scale. For example, test takers might be placed into *Proficient* and *Nonproficient* categories or into levels of certification, achievement, or competence. The performance

reporting for these types of tests would be oriented more to classification results and less to the properties of an entire range of scale scores. Cut scores and classifications would have increased importance, such that representing test performance relative to cut scores involves a classification problem with unique implications for measurement, CSEMs (Lord, 1985), and other properties of the classifications.

Some proposed procedures for cut scores are adaptations of what was described in the *Reporting Scale Basics* section. For tests with a long range of reported scores, Dorans (2002) recommended scale scores centered at the midpoint of the scale, with a working range that extends beyond the reporting range and with scale score units that do not exceed the number of possible raw score points. For tests with a small number of cut scores, Dorans et al. (2010) recommended that scales be established that are centered near the cut score(s), with score points not exceeding the number of raw score points, and with working ranges wider than reporting ranges, all of which would accommodate shifts in test difficulty and potential additions of new cut scores. For both types of tests, the score scale should be regarded as infrastructure likely to require repair and corrective action. Recommendations are for test assembly and scaling procedures that focus on the cut scores of interest. Also, CSEMs that are equal across the entire range of a test's scores might be less important than ensuring that the CSEMs are small near a cut score of interest (Lord, 1985).

Other procedures for establishing scale scores for tests with cut scores are content based and not data, sample, or population based as described in Table 11.1. The most well known of these are standard-setting procedures (Hambleton & Pitoniak, 2006; Ferrara et al., this volume), which begin with a statement of what competent test takers know and are able to do and then search for potential cut scores judged to reflect these statements. Various standard-setting methods can be used to ask a set of judges to work with or produce statements for what they think test takers should be able to do at a specific performance level and then identify through judgmental processes the test questions and performance that correspond to these statements. Although the basis of standard-setting procedures is primarily judgment and test content, the process is often supplemented with test performance data, such as item and test performance, and results indicative of the impact of cut score recommendations on scale score distributions.

Scales for Test Batteries and Composites

For a battery of tests covering different content areas, scales might be established using the same scaling methods with all tests in the battery, as taken by the same test-taker group. The resulting scales for each test in the battery can facilitate similar interpretations for scores in each of the tests, such that test takers' strengths and weaknesses are revealed in their scores on the battery. One of the earliest examples of battery scaling was proposed by Kelley (1914, 1923) for establishing comparable units (i.e., same means and standard deviations in the reference population) for the handwriting scoring method proposed by Thorndike (see the *Scaling Perspectives* section) and another scoring method. From a discussion of alternative proposals for establishing comparable scores from the two scoring

methods of handwriting, Kelley proposed establishing comparable measures as standard scores similar to Equation 9. More recent examples of battery scaling are the recentered SAT (Dorans, 2002), where the SAT Math and Verbal scales were both established based on normal distributions with means of 500 and standard deviations of 110. The 2016 redesigned SAT also targeted similar Math and Evidence-Based Reading and Writing (ERW) Section score means and standard deviations (Kim et al., 2016). The ACT rescaling targeted means of 18 with constant SEMs (Brennan, 1989), but different scale score standard deviations. The scale scores for each of these batteries can encourage different types of interpretations, such that test batteries for normalized scales might facilitate references to percentiles of the normal distributions, those established with similar means and standard deviations might facilitate references to standard deviation units, and those established based on stabilized SEMs might encourage SEM-based score distinctions rather than distinctions based on standard deviation units (Kolen, 2006).

Composite scores might be formed based on linear combinations of the scores on each test in the battery. For example, the Total score of the 2016 redesigned SAT is the sum of the Math and Evidence-Based Reading and Writing Section scores, and the ACT Composite score is the average of the math, science, English/writing, and reading test scores. Both examples are reflections of nominally weighted scores. Other weighting procedures such as those based on effective weights and the proportional contribution of test score variances to the composites can produce different composite score scales, to the extent that the test scales have different standard deviations and the tests have different correlations with each other (Brennan, 1989; Kolen, 2006). The scaling characteristics of batteries and composites of multiple tests are affected by the scaling results for the individual tests, the scale score standard deviations of the individual tests, and the intercorrelations among the scale scores of the tests. Because scales are usually set and maintained on individual tests, the scale score characteristics for batteries and composites are more difficult to establish and maintain. The *Battery Scaling and Composites* section provides further discussion and an example of these issues.

Auxiliary Scales

After one or more primary scales are set, additional procedures and methods can be used to convey enhanced interpretations of the primary scale scores. For auxiliary scales, the goal is less about building interpretations directly into the scales themselves and more about supplementing the primary scales with additional interpretive information. Two categories of interest are based on normative information and on content.

Auxiliary Scales Based on Norms

Norms refer to the statistical information that could be provided and used to describe performance associated with different scale scores, such as the scale score mean, standard deviation, and percentile ranks (Equation 11 applied to $sc(Y)$) for one or more test-taker groups. Norms might be determined for groups sampled from particular populations. Sometimes the group used to establish norms is the same sample from

population P used to establish the scale scores, in which case the norms of $sc(Y)$ are consistent with the statistics (i.e., means, standard deviations) directly established in the distribution of $sc(Y)$. Previous volumes of *Educational Measurement* have listed options for norms groups and reporting (Angoff, 1971; Flanagan, 1951; Kolen, 2006; Petersen et al., 1989), presented here with an update.

- National norms: Describe test performance from nationally representative test takers at the age or educational level for which the test was designed.
- Local norms: Describe test performance from test takers from specific educational or geographic units.
- User norms: Describe test performance from test takers who take the test during a given time period.
- Convenience norms: Describe test performance from test takers who are available at the time a test is constructed.
- Group-level norms: Describe aggregated test performance from groups, such as average performance for schools, districts, or states.
- Item-level norms: Describe the performance on specific items for a norm group.
- Skill-level norms: Describe the performance on sets of items measuring a particular skill for a norm group.
- Growth norms: Describe test performance for a group of students (e.g., K–8) that takes at least two tests at different grades, usually where separate sets of norms on the second test are reported for subgroups of students obtaining different scale scores on the earlier test (Betebenner, 2009; Castellano & Ho, 2013).

Typically, national norms that describe test performance for nationally representative test takers need the most technically elaborate norming studies. National norming studies require that the national population of interest is precisely defined in terms of students, schools, etc. A sampling plan is developed to include test takers for whom norms will produce accurate estimates of the population norms. National norming studies may use several sampling approaches to increase the representativeness of the test takers and resulting norms, including simple random sampling, random sampling within defined strata of student characteristics (e.g., geographic region, school type), cluster sampling of test-taker groups (e.g., schools), or systematic random sampling that reflects the ordering of a variable related to the test scores of interest in the norms study. Often, national norming studies use a combination of sampling and statistical weighting procedures to obtain schools and test takers and increase their national representativeness. Approaches to national norming studies by the National Assessment of Educational Progress (NAEP, <https://nces.ed.gov>) and by ACT were compared in terms of their methods, complexity, and practical challenges by Kolen (2006). Other types of norms (user, convenience, etc.) might be produced using less elaborate procedures (e.g., test takers available at the time the norms are produced). Such norming groups may be less stable and more likely to change (Petersen et al., 1989).

Auxiliary Scales Based on Content

Some auxiliary scales are attempts to supplement primary score scales with content-based descriptions of specific scale scores. One example involves item mapping, or finding a set of items that represents various scale score points. In the case of NAEP, these items “help to illustrate what students know and can do in NAEP subject areas by positioning descriptions of individual assessment items along the NAEP scale at each grade level” (<https://nces.ed.gov>). For item mapping, the scores of individual items are regressed on scale scores using logistic regression or an IRT model. Then a response probability is adopted to divide test takers into groups with “higher” and “lower” probabilities of success on test items. The actual values used to divide the higher and lower probability groups could be .5/.5, .8/.2, or other probabilities for dichotomously scored items. In NAEP, these groups are divided using probabilities based on 74% for MC items and 65% for CR items. These response probability values were obtained based on work by Huynh (1998), who showed that these values correspond to the maximum IRT-based information provided by a correct response to MC or CR items. The result of item mapping is a list of test questions that represents various scale score points, with content-based descriptions that take the form of phrases that describe what test takers can do correctly, in general terms rather than for one or more specific items (Kolen, 2006; Kolen & Brennan, 2014).

Scale anchoring is another type of content-based auxiliary scale, with the goal of providing general statements about what students obtaining different scale scores would know and are able to do (Kolen, 2006). Item maps are a first step to scale anchoring. Additional steps involve choosing a set of scale score points across the score scale range, where items that map at or near these points are chosen to represent these points. Then subject matter experts review the items mapping near each point and produce statements presenting the skills of test takers scoring at these points, assuming test takers also know and are able to do all of the skills in the statements at or below the given score level. Scale anchoring has been used in NAEP (Allen et al., 1999) and in ACT (2001). The Binet–Simon (Binet & Simon, 1916) test described in the *Scaling Perspectives* section can also be viewed as a test with sets of items specifically anchored to different scale scores. Item-mapping and scale-anchoring procedures have raised questions about whether their outcomes actually facilitate score interpretations (Forsyth, 1991). The content-based labeling produced with scale anchoring appears to be the result of overly subjective and confusing judgmental processes (Pellegrino et al., 1999) that may not be supported by statistical analyses (Haberman, Sinharay, & Lee, 2011; Thurstone, 1925) or by items that are sufficiently discriminating (Dorans, 2018).

Summary of Primary and Auxiliary Scales

Reporting scales should be viewed as infrastructure in need of monitoring and possible repair (Dorans, 2002). For the primary scaling approaches reviewed in the *Scaling* section, potentially important checks for primary scales include checks for excessive accumulations at the top, middle, or bottom parts of the scale; checks that the score means do not excessively depart from the midpoint of the scale; and other checks that

properties like distribution shapes, CSEM values, etc., are established and maintained as intended. As described in the *Scales for Exams with Cut Scores* section, primary scales can also be established and monitored in terms of cut scores of interest, including distributions resulting from pass/fail decisions and CSEMs at or near those cut scores. These possible checks suggest different types of monitoring for primary scales established in different testing contexts. In admissions testing where testing companies develop tests for users to employ different parts of a relatively long scale, monitoring should focus on the entire score range. In K–12 and certification and licensure programs, the focus is primarily on classifications at one or very few cut points, and the monitoring of scale scores would likely focus on pass/fail distributions, classification accuracies, and CSEMs at specific cut scores. Scale maintenance for large-scale survey assessments like NAEP might focus on the estimation of population and subpopulation scale score distributions and classifications over time. In each of these contexts, scales established in special studies should be checked to ensure that their results will accurately apply beyond these scaling studies and to the intended testing population. The auxiliary scales that might be established and reported for a particular testing program should also be monitored for potential changes due to testing population changes and other potential changes in content-based descriptions. The equating procedures described in the *Equating* section are needed for preserving established primary and auxiliary scales.

EQUATING

For a large-scale testing program, the establishment of a reporting scale as described in the *Scaling* section usually comprises only an initial step for a single test form (Y). Usually Y is the first of many test forms to be developed and administered. Additional work is needed to develop, administer, and report scores for the subsequently developed test forms for that testing program. Equating is used to address unintended difficulty differences in the scores of these alternate forms, so that the previously established reporting scale is maintained (Holland & Dorans, 2006; Kolen & Brennan, 2014). This section describes equating in an updated version of the discussion from Holland and Dorans (2006), which itself was a continuation of previous discussions from Petersen et al. (1989), Angoff (1971), Flanagan (1951), and others.

The *Distinguishing Equating From Other Forms of Linking* section describes some general concepts about equating and its distinctions from other types of linking based on goals, interpretations, and computational procedures. Equating history and requirements are covered in the *History of Equating* and *Equating Requirements* sections. The data collection designs used in different types of equating are described in the *Data Collection Designs* section. The final section covers *Methodological Implementations of Equating*. The different equating approaches covered in these discussions are summarized in Table 11.2 in terms of the equating definition, method(s), assumptions, and likely data collection designs.

Table 11.2 Summary of Types of Equating

Definition of Equating	Method(s)	Assumptions	Data Collection Design(s)
Align the means and standard deviations of observed scores of alternate test forms (common population)	Linear scale alignment	Tests measure the same construct; population invariance; equity	Randomly equivalent groups; single groups; counterbalanced
Align the distributions of observed scores of alternate test forms (common population)	Equipercentile		
Align the means, standard deviations, and/or distributions of observed scores of alternate test forms (different populations)	Tucker, Braun–Holland; frequency estimation/poststratification	Tests measure the same construct; population invariance in the conditional test-given-anchor distributions; equity	NEAT
	Chained linear; chained equipercentile	Tests measure the same construct; population invariance in the two chained functions; equity	NEAT
	Levine observed score; modified frequency estimation	Tests and anchors are congeneric; population invariance in the conditional Test-given- $\hat{\tau}_A$ distributions; equity	NEAT
	IRT observed score	Items measure the same construct; the IRT model fits the data; IRT parameter estimates for all items are on the same scale and invariant across administrations; equity	NEAT (this method is also used with randomly equivalent groups, single group, and counterbalanced)
Align the means and standard deviations of true scores of alternate test forms (different populations)	Levine true score	Tests and anchors are congeneric; true score results can be applied to observed scores; equity	NEAT
Link true scores via an expected test-given-ability relationship based on an IRT model	IRT true score	Items measure the same construct; the IRT model fits the data; IRT parameter estimates for all items are on the same scale and invariant across administrations; true score results can be applied to observed scores; equity	Randomly equivalent groups; single group; counterbalanced; NEAT
Score test takers using item parameter estimates obtained and linked in previous administrations	IRT ability estimation	Items measure the same construct; the IRT model fits the data from the previous calibrations; parameter estimates are invariant across administrations with minimal context and order effects; equity	Pre-equating

Note. IRT = item response theory; NEAT = nonequivalent groups with anchor test.

Distinguishing Equating From Other Forms of Linking

Equating is one type of linking procedure for transforming the scores of one test form to another (Holland, 2007; Holland & Dorans, 2006; Pommerich, 2016). Specifically, equating adjusts the scores of test forms constructed and administered in the same way for unintended difficulty differences. Within a linking hierarchy, equating is the strongest among several types of linking, distinguished in terms of interpretations and computational methods (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Mislevy, 1992; Petersen et al., 1989).

Different categories of linking can be used to achieve different goals and support particular interpretations of the linked scores. When equatings of alternate forms of the same test are produced (X and Y), the expected result is interchangeability in the reported scores, meaning that when form X is equated to form Y , the resulting X -to- Y scores can be used in place of Y 's scores for any purpose, as if those scores came from the same test (Dorans, 2013; Pommerich, 2016). For other types of linking involving different tests (i.e., X and Y are different tests), weaker and more limited results are expected. One alternative goal is comparable scales (Angoff, 1971; Kelley, 1923; Pommerich, 2016), which is one component of interchangeability that can be defined as the result of a symmetric alignment of scale score distributions that facilitates comparisons of the resulting scores. In the absence of interchangeability, comparability is group dependent (see the *Linking* section).

Another goal of linking is “best” prediction, which is an asymmetric conversion from the scores and scales of one test to another, such that prediction error is minimized. Predictions are distinguished from interchangeability and comparability in that predictions are often group dependent and always asymmetric (i.e., for given X scores, the best prediction from X to Y does not equal the best prediction from Y to X).

Scale-Aligning Computations

Scale-aligning methods are one type of computation used in test linking, including equating (Holland & Dorans, 2006). These methods can be described as score transformations that result in the scores of two test forms having the same distribution (Kelley, 1923). For example, consider test Y as a form for which a scale score transformation has been established, $sc(Y)$, and X is another test form whose scores are intended to be transformed to align to the same scale. The transformation of X to the scale of Y may be achieved through matching the marginal distributions of these forms obtained from test takers representing target population T (Holland & Dorans, 2006, p. 202, defined T as the source of the equating data and also presented other perspectives on target populations in equating). An equipercentile X -to- Y transformation produces Y scores with percentile ranks that equal those of every X score for X and Y test takers from the target population T . The equipercentile linking is produced through computing percentile ranks at each score of X (Equation 11 for score x on the distribution of X , $F_T(x)$) and obtaining the inverse of Equation 11 on the distribution of Y (i.e., $G_T^{-1}[y]$) in T ,

$$e_{Y,T}(x) = G_T^{-1}[F_T(x)]. \quad (15)$$

Equation 15 produces a symmetric X -to- Y conversion that equals the inverse of the Y -to- X conversion at the same x scores, $e_{X,T}^{-1}(y)$ at x . The X -to- Y scores have a distribution on X that approximately equals the distribution of the Y scores in population T . Simpler versions of Equation 15 can also be produced by using a version of Equation 9 to solve for X -to- Y scores that match the mean and standard deviation of Y but not other aspects of Y 's distribution,

$$l_{Y,T}(x) = \mu_{Y,T} + (x - \mu_{X,T}) \frac{\sigma_{Y,T}}{\sigma_{X,T}}, \quad (16)$$

where $\mu_{X,T}$ and $\mu_{Y,T}$ denote the population means of X and Y in T , and $\sigma_{X,T}$ and $\sigma_{Y,T}$ denote the population standard deviations. Equations 15 and 16 establish comparability through symmetric transformations that match the score distributions (Angoff, 1971; Holland & Dorans, 2006; Kelley, 1923; Pommerich, 2016). Once the results are obtained from either Equation 15 or Equation 16, X -to- Y scores can be transformed to the reporting scale by applying one of the $sc(Y)$ scale transformations described in the *Scaling* section.

Projecting and Predicting Computations

Projecting and predicting are types of linking procedures that differ from scale-aligning approaches like equating. Rather than aligning the scores of form X and form Y to have comparable score distributions for a particular group of test takers, these methods are used for the goal of producing an asymmetric, best prediction of scores or a projection of score distributions for one test to those of another. Predictions and projections can also be produced for the IRT-based thetas from different tests (Holland & Hoskens, 2003; Thissen et al., 2015). One example of a projected distribution is the percentile rank of $Y = y$ for a subgroup of test takers obtaining a specific score on $X (= x)$, $F_T(y|x)$. Another example is the prediction of the means of Y given X using a linear X -to- Y regression,

$$reg_{Y,T}(x) = \mu_{Y,T} + (x - \mu_{X,T}) \frac{\sigma_{Y,T}}{\sigma_{X,T}} \rho_{XY,T}, \quad (17)$$

where $\rho_{XY,T}$ is the XY correlation. The prediction shown in Equation 17 is a linear regression that provides an expectation of Y given X that minimizes prediction error, the variance of $Y - reg_{Y,T}(x)$, an essential supplement to Equation 17's point estimates (Thissen et al., 2015). Unlike some computational methods for scale aligning, projecting and predicting functions require data collection designs where test takers take both X and Y . Also unlike scale-aligning functions, predicting functions are not symmetric, meaning that $reg_{X,T}^{-1}(y)$ at x is not equal to $reg_{Y,T}(x)$.

Some results are of special interest for situations in which there is a perfect one-to-one relationship in X and Y . For linear functions, this situation means that the correlation between X and Y is 1 and the X -to- Y regression in Equation 17 produces the same results as the X -to- Y scale alignment in Equation 16. This result has several uses. One use is defining a situation in which a regression can be used to align scales. For example, consider the second half of Equation 6, $\sum_i w_i \Pr(V_i = 1 | \theta, a_i, b_i, c_i) = \tau_Y(\theta)$,

which can be described as a regression of IRT-implied true scores on θ (Lord & Novick, 1968, p. 386), such that these true scores have a perfect and functional (not statistical) nonlinear relationship with θ under the IRT model. When same-scale IRT parameter estimates are available for the items of tests X and Y , the IRT-based true score regressions of X and Y can be combined to produce a scale alignment of the true scores of X and Y , $\tau_Y[\tau_X^{-1}(\theta)]$, where $\tau_X^{-1}(\theta)$ denotes the inverse of the θ -to- τ_X function (i.e., τ_X -to- θ , Lord, 1980). In addition to these IRT applications, perfectly correlated X and Y true scores might also be assumed and used to produce scale alignments of τ_X and τ_Y based on classical congeneric theory (see the *Equating Requirements* section; Kolen & Brennan, 2014; Levine, 1955). Another use of the relationship of Equations 16 and 17 is to describe the strength of a linking of tests with scores that are not perfectly correlated. For example, in concordances of different tests (see the *Concordances* section), the correlation has been used as a measure to diagnose the strength of the resulting concordances (see the *Correlations and Prediction Error* section), such as by describing the extent to which a scale alignment might be used to produce score predictions that approximate the accuracy of projection or prediction functions (Dorans, 1999; Moses, 2014a).

History of Equating

The origins of equating date to the early 20th century, when an intelligence test known as the Army Alpha was developed for selection into the military for World War I (Yoa-kum & Yerkes, 1920). Alternate test forms were developed by randomly assigning items to the forms, to prevent cheating and to establish forms of approximately equal difficulty. The difficulties of the alternate forms were subjected to empirical evaluations, and for the subset of forms surviving these evaluations, the unadjusted observed scores were treated as interchangeable. The linear linking function (Equation 16) was recommended over the nonsymmetric regression function (Equation 17) for establishing comparable measures on the basis of approximating scale score distributions (Otis, 1922; Thorndike, 1922). Kelley (1923) provided more discussions of linear linking functions and an introduction of the equipercentile method (Equation 15). Nevertheless, “the need, or at least the desire, to equate scores on alternate forms of the same test probably arose decades after the invention of scaling methods and of the two standard methods for equating—the linear and equipercentile methods” (Holland & Dorans, 2006, p. 196).

Equating was described in general and in practice for the Cooperative Achievement Tests by Flanagan (1939, 1951). These discussions covered equating for the data collection designs currently referred to as the single group, randomly equivalent groups, and counterbalanced designs (see the *Data Collection Designs* section). Other emphases were on form construction and linear and equipercentile methods. Test equating for nonequivalent groups with an anchor test (the NEAT design, see the *Data Collection Designs* section) has been traced to the SAT, which administered two editions in 1938 and employed anchor equating in 1941. For this SAT application, a method attributed to Tucker and eventually named Tucker equating (Angoff, 1971) was used to implement a linear scale alignment (Equation 16) with regression-based predictions of the X and Y means and standard deviations for a hypothetical target group (Gulliksen, 1950; Holland & Dorans, 2006; Lord, 1950).

Since the early SAT equatings, additional discussions and uses have been provided, including Flanagan's (1951) comprehensive treatment of test development methods in equating, computational methods, data collection designs, a definition of comparability in equating, and, notably, a warning about likely population sensitivities and group dependences in linking results (p. 748). Lord (1950) provided a statistical overview of linear equating procedures and standard errors for various data collection designs. Angoff (1971) provided another comprehensive discussion of equating, computational methods and data collection designs, which, in contrast with Flanagan (1951), emphasized the goal of equating to produce results that are relatively group invariant (p. 563). Angoff also described situations where a linear equating function approximates the equipercentile function (i.e., when the shapes of the X and Y distributions are similar) and recommended linear functions for these situations (p. 564). Angoff's (1971) chapter has been a foundational and influential reference for decades of practice and literature on equating, scaling, and linking.

Updates to the statistical aspects of equating were provided by Holland and Rubin (1982), including a discussion of the mathematical properties of commonly used equating methods (Braun & Holland, 1982) and section pre-equating (Holland & Wightman, 1982). Interspersed in this history, proposals were made to describe and implement equating based on measurement theory, including classical true score theory (Levine, 1955), IRT (Lord, 1980; Lord & Wingersky, 1984), and other works focusing on theoretical properties such as equity of equating results with respect to true scores (Hanson, 1991; Morris, 1982) and IRT (van der Linden, 2011). The IRT developments were especially important for providing alternative theoretical and practical implementations for equating, including proposals for item-level pre-equating and IRT-based adaptive testing (Lord, 1980). Discussions and updates have continued, including the chapters in the third and fourth editions of *Educational Measurement* (Holland & Dorans, 2006; Petersen et al., 1989), subpopulation invariance evaluations (Angoff & Cowell, 1986; Dorans & Holland, 2000; Harris & Kolen, 1986), several texts (Dorans et al., 2007; Holland & Dorans, 2006; Kolen & Brennan,

2014; Linn, 1993; Livingston, 2004; Mislevy, 1992; von Davier, 2011; von Davier et al., 2004a), and software packages (Brennan, 2004; Brennan et al., 2009; Gonzalez & Wiberg, 2017).

In practice, equating has been implemented using different data collection designs for different testing programs, including programs that equate with randomly equivalent groups designs (ACT and the 2016 redesigned SAT) and others that use anchor tests to equate through nonequivalent administration groups designs (the SAT prior to the 2016 redesign). Other programs use designs suited for small samples (Puhan et al., 2009), and for pre-equating (CLEP or College-Level Examination Program, Gao et al., 2012). These implementations of test equating are summarized in Table 11.2.

Equating Requirements

In discussions of linking frameworks (Holland & Dorans, 2006), equating is described as the strongest form of linking and is subject to the strictest requirements. These requirements were listed by Holland and Dorans (2006) as a culmination of several previous discussions of equating (Angoff, 1971; Dorans & Holland, 2000; Flanagan, 1951; Kolen & Brennan, 2014; Linn, 1993; Mislevy, 1992; Petersen et al., 1989). The equating requirements are summarized in this section, including updates and clarifications from Brennan (2010), Dorans and Walker (2007), and Kolen (2007).

An equating can be used to align the scales of X and Y , such that X -to- Y and Y scores can be used interchangeably, thereby maintaining a testing program's reporting scale. For an X -to- Y transformation to be considered an equating, the X and Y test forms and the X -to- Y transformation must satisfy the following requirements:

- a. The equal construct requirement: Test forms X and Y should measure the same constructs.
- b. The equal conditions of measurement requirement: The same administration and measurement conditions should be used for X and Y .
- c. The equal and high reliability requirement: The reliabilities of X and Y , $rel_{X,T}$ and $rel_{Y,T}$, should be equal and high.
- d. The equity requirement: The X -to- Y equating should make it a matter of indifference whether test takers take X or Y , for test takers at every given ability level.
- e. The (sub)population invariance requirement: The X -to- Y equating should be population invariant, in that the choice of (sub)population used to estimate it does not matter.
- f. The symmetry requirement: The X -to- Y equating function should be the inverse of the Y -to- X equating function.

Discussions of the equating requirements describe them as aspirational concepts (Brennan, 2010; Dorans & Moses, 2023) that provide an intuitive theory of test equating (Holland & Dorans, 2006). These requirements are not likely to be perfectly met in practice, sometimes prompting criticisms for being vague, impractical, stringent,

unnecessary or impossible (Dorans & Holland, 2000; Livingston, 2004; Lord, 1980). Nevertheless, the equating requirements specify the goals of the test development, administration, scoring, and linking processes that are useful for producing interchangeable scores that maintain the reporting scales of a testing program for a target population of test takers. As described in *Linking*, the equating requirements are also useful for characterizing nonequating types of linking.

Equating requires that the test forms being equated are assembled according to the same content and statistical specifications, such that they exhibit the same measurement properties (Requirement a). Equating Requirement b emphasizes consistency in administration conditions for tests that are equated (Kolen, 2007). In describing conditions of measurement, Kolen (2007) listed aspects under the control of the test developer such as that tests *X* and *Y* are administered with the same instructions, layout, timing, scoring procedures, aids, and modes (computer or paper–pencil). Other measurement conditions that are not under the direct control of the test developer can also affect the quality of equating outcomes, including stakes for test performance, reasons test takers take the test, and type of test preparation activities.

Requirement c, that the test forms are highly and equally reliable, has received ongoing and updated clarifications. The equal reliability aspect of this requirement reflects earlier recommendations, meaning that the test forms being equated must have the same measurement precision. The requirement that forms' reliabilities be high is a more recent addition, which conveys that equating requires scores that are precise and close reflections of their true scores, that support high correlations and prediction accuracy, and that make subpopulation invariance (Requirement e) and equity (Requirement d) more likely (Brennan, 2010; Dorans & Walker, 2007; Flanagan, 1951; Holland & Hoskens, 2003; Kolen, 2004; Lord, 1980).

As stated in the *Scaling* section, many issues concerning reliability, true score variance, and error variance are, at best, imperfectly understood. They are also increasingly complex when reliability and equating are considered in greater detail and in relation to other equating requirements. The reliability requirement depends on consistently following test development procedures and specifications for test length, content, items, administration, and scoring procedures (Requirements a and b). Test forms developed from different sets of items that meet the same specifications could be assumed to be parallel in the classical test theory sense (equal reliabilities, equal true scores, and errors with equal variances). In practice, assembled forms can be imperfectly or nominally parallel (Lord & Novick, 1968) and might have unintended differences, such as differences in their mean scores attributable to their unique samples of items. Equating can be described as the application of scale-aligning adjustments that address errors attributable to forms' item samples, errors that are realized in fixed difficulty differences of the specific form(s) being equated (Li, 2023; Moses & Kim, 2015). Equating does not address the reduction of relative errors (i.e., interactions of test takers and items, as well as other unspecified sources of error; Li, 2023).¹

Forms' error variances can also be considered conditional on test-taker ability, variances that are assumed to be equal according to Equity Requirement d. One proposal based on equity applies conditional, ability-specific scale-aligning adjustments (van der Linden, 2011, 2013), which could be interpreted as an attempt to adjust scores for IRT-implied relative errors, with the forms' items and parameters being treated as fixed. As described in the *Local, Ability-Specific Test Linking* section, these local adjustments likely violate other equating requirements concerning consistency in scoring rules for a given form (Requirement b) and subpopulation invariance (Requirement e).

Different score models and associated reliability measures have equating implications and provide unique representations of reliability and error, including model-implied error variances based on fixed items (commonly used IRT models), score models that treat items as random samples, and more complex score models that account for errors across a range of admissible measurement conditions. These issues illustrate the lack of a single framework for measurement error and other types of equating error, the need for greater specificity in reliability and equating, and the importance of test form reliabilities that are high and equal in some sense for reducing the impact of these ambiguities on equating results.

Equity Requirement d states that the conditional distributions of Y and X -to- Y equated scores are equal for test takers at specific values of a latent ability based on a particular measurement model (Brennan, 2010; Hanson, 1991; Lord, 1980; Morris, 1982; van der Linden, 2011). Equity was originally defined in terms of IRT models, θ s, and θ -conditional frequency distributions (Lord, 1980). Equity is commonly expressed based on θ -conditional cumulative distributions (Kolen & Brennan, 2014; van der Linden, 2011),

$$F_{\theta}[e_{Y,T}(x)] = G_{\theta}[y] \quad \text{for all } \theta, \quad (18)$$

where the θ subscript indicates conditioning on θ .

Equity discussions based on classical test theory have focused on relationships among equating functions of observed scores, true scores, and equating requirements (Holland & Dorans, 2006; Hanson, 1991; Kolen & Brennan, 2014; Morris, 1982). Holland and Dorans (2006) attributed an equity theorem to Hanson (1991) and synthesized several equating requirements. To summarize, assume that X and Y measure the same construct and follow a congeneric model (i.e., their true scores are perfectly and linearly related) and the functional relationship of the true scores, τ_X and τ_Y , can be expressed as a scale aligning function,

$$\tau_Y = l_{\tau_Y}(\tau_X) = \mu_{Y,T} + (\tau_X - \mu_{X,T}) \frac{\sigma_{Y,T}}{\sigma_{X,T}} \frac{\sqrt{rel_{Y,T}}}{\sqrt{rel_{X,T}}}. \quad (19)$$

Substituting observed X scores into true score conversions like Equation 19, $l_{\tau_Y}(x)$ satisfies first-order equity where the conditional mean of $l_{\tau_Y}(X)$ given $l_{\tau_Y}(\tau_X) = \tau_Y$ equals the conditional mean of Y given τ_Y for all τ s. Hanson (1991) showed this first-order equity result in Levine true score equating for observed scores (see the *True Scores* section) with the nonequivalent groups with anchor test (NEAT) design (see the *Data Collection Designs* section). Holland and Dorans (2006) additionally stated that when X and Y are equally reliable, $l_{\tau_Y}(x)$ equals $l_{Y,T}(x)$ in Equation 16 and satisfies second-order equity where the conditional variance of $l_{\tau_Y}(X)$ given $l_{\tau_Y}(\tau_X) = \tau_Y$ equals the conditional variance of Y given τ_Y for all τ s. Brennan (2010) provided a related discussion that considered quadratically represented curvilinear equating functions, showing that curvilinear equating functions are more likely to satisfy first- and second-order equity if the reliabilities of X and Y are not only equal, but also high (nearly linear equating functions are also helpful, but less realistic). Altogether, these results connect several equating requirements, such that X and Y are assumed to measure the same constructs (Requirement a), are administered under the same conditions (Requirement b), with scores that have equal and high reliabilities (Requirement c), and where the equating is a symmetric scale alignment of the forms' observed scores (Requirement f) that approximates the functional relationship of their true scores and achieves first- and second-order equity (Requirement d) (Holland & Dorans, 2006). The results represent one of the most general theories of equating available, including all but one equating requirement (i.e., the subpopulation invariance requirement is out of scope, since the results are based on a fixed target population, T ; Holland & Dorans, 2006).

Equating Requirement e emphasizes that the same X -to- Y linking results should be obtained if estimated in T and in subpopulation T_g ,

$$e_{Y,T_g}(x) = e_{Y,T}(x) \quad \text{for each } T_g. \quad (20)$$

The subpopulation invariance requirement is more likely met when X and Y are constructed to be similar and of high reliability (Dorans & Holland, 2000; Flanagan, 1951; Kolen, 2004). The invariance requirement of equating has been especially important for use in the empirical evaluations of nonequating types of linking (see the *Linking* section) and also for linkings intended to be used as equatings. Indices for quantifying the extent to which equating functions vary by subgroup are summarized in the *Subpopulation Invariance Evaluations* section.

Requirement f indicates that asymmetric prediction and projection functions cannot be used as equating functions because asymmetric functions mean that it is not a matter of indifference which form test takers take. Symmetric functions like Equations 15 and 16 can be described as a computational realization of comparable scales (i.e., comparable scale score distributions; Holland & Dorans, 2006; Kelley, 1923).

Data Collection Designs

Data collection designs for test equating are summarized in Table 11.1 and described in detail in several texts (Angoff, 1971; Dorans, 2018; Dorans et al., 2010; Holland & Dorans, 2006; Kolen & Brennan, 2014; Kolen, 2007; Petersen et al., 1989; von Davier et al., 2004a). Each design is an approach to obtaining test data such that an equating of the forms' scores will estimate differences in test difficulty and other characteristics, where the estimates (a) control for the abilities of the test-taker groups taking both X and Y and (b) reflect target population T . Successful data collection designs are essential for obtaining data from which accurate equating and other scale alignment and projection functions can be estimated.

The simplest test equating design shown in Table 11.1 is the single group design, where one sample of test takers from population $P (= T)$ takes both X and Y . This design can use relatively small samples to produce statistics that accurately reflect their population values. The necessary conditions for this design are that it must be realistic for test takers to take both tests and that the resulting data are not adversely affected by timing issues, testing fatigue, learning, or order effects. Scale-aligning, predicting, and projecting methods are all possible with the single group design, but satisfying the necessary conditions can be challenging. One advantage of this design is that a correlation between the two tests can be estimated to enable a direct check on Requirements a and c, provided that Requirement b is also met.

Another design shown in Table 11.1 is the randomly equivalent groups design, where two independent samples of test takers from population $P (= T)$ are randomly assigned to take X or Y . In large-scale testing, the random assignment is usually implemented with spiraling, where X and Y are alternated in their delivery, resulting in systematically assigned test forms to the test-taker groups. Groups end up being more equivalent than would be expected with simple random sampling (von Davier et al., 2004a). Because independent samples are used, the randomly equivalent groups design requires larger sample sizes to produce statistics for test forms X and Y with the same precision as obtained with the single group design. When test-taker sample sizes are large and the test forms can be reused and readministered without security problems, the randomly equivalent groups design has advantages in that it avoids order effects that can arise with the single group design. Scale-aligning methods are used with the randomly equivalent groups design.

In the counterbalanced design shown in Table 11.1, independent samples are drawn from population $P (= T)$, one where test takers take form X and then Y , and another where test takers take Y and then X . This implementation combines the single group and randomly equivalent groups designs, allowing for order effects to be estimated based on the differences in test linking results for data reflecting each testing order. To the extent that order effects are observed, various choices can be made for linking forms X and Y with the counterbalanced data (von Davier et al., 2004a), including the use of all available data to represent both testing orders, the use of only subsets of the data that do not reflect order effects (where some test takers take X first and other test takers take

Y first), and other weighted combinations of these. Usually a counterbalanced design is implemented in a special study where the estimation of order effects is of interest, rather than in a typical test administration where testing orders are fixed. As with the single group design, Requirements a and c can be checked, as can Requirement b.

Another data collection design shown in Table 11.1 is less direct, where X is linked to Y through scores on an anchor test, A . This is the NEAT design (Holland & Dorans, 2006), which is also referred to as the common-item nonequivalent groups design (Kolen & Brennan, 2014). Samples from two different populations take X or Y , and both samples take A . Because the samples represent different populations, estimates of population differences are needed for the X -to- Y linking, and the groups' performance on A provides these estimates. The NEAT design contains two single group designs. In practice, A can be a set of items common to X and Y (an internal anchor) or a test or set of items that is separate from X and Y (an external anchor). The quality of the final linking results based on internal or external anchors reflects the extent to which A is representative of X and Y . Accordingly, data collections and procedures are encouraged that strengthen the anchor test (Dorans et al., 2011; Holland & Dorans, 2006).

The NEAT design is more complex than the single group, randomly equivalent groups, and counterbalanced designs. Specifically, the NEAT design requires invariance assumptions for estimating test-taker group performance on the test that test takers do not take (assumptions about the unobserved Sample 1 from population P on Y and about the unobserved Sample 2 from population Q on X). Invariance assumptions must be made to produce an equating that applies to the target population. Two major approaches and their associated assumptions involve chaining through A and estimating X and Y distributions by projecting from A for a synthetic population defined as a combination of P and Q (described in the *Approaches to NEAT Equating* section).

An approach similar to the NEAT design involves assembling X using items from an item pool that have IRT parameter estimates on the same scale, X_A , along with other newly administered items, X_{New} . In this common item equating to a calibrated pool design (Table 11.1, Kolen & Brennan, 2014), the X_A items are used as an anchor to obtain scaled IRT parameter estimates for the X_{New} items that expand the item pool. Then, as described in the *IRT* and *Distinguishing Equating From Other Forms of Linking* sections, an IRT conversion to the θ scale, to Y , or to any other form that might be assembled from the item pool, is produced for the scores of the complete X form (X_A and X_{New}). Parameter estimates for the item pool and the equating results should all apply to population P ($= T$). Common item equating to a calibrated pool is more flexible than the NEAT designs with respect to the source of the anchor items, but it requires an IRT implementation and the meeting of IRT assumptions (Kolen & Brennan, 2014).

The final rows in Table 11.1 present two pre-equating designs for linking X to Y using data from previously administered sections or individual items to equate X before it is ever administered as an intact test. Section pre-equating involves the administration of a complete old form, Y , composed of two mutually exclusive sections (Y_1 and Y_2) and one of two mutually exclusive sections of new form X (X_1 or X_2) to randomly

equivalent samples of population $P (= T)$. The statistics for the complete X form could be estimated using data from the administration of the X_1 or X_2 sections, and these estimated statistics could be used to link X to Y (Holland & Wightman, 1982). The use of two sections is the simplest case for this design. In their discussion of section pre-equating, Petersen et al. (1989) presented implementations with three and four sections of X and Y administered to three and six random samples of P .

Pre-equating can also be implemented in individual items rather than test sections. The final row of Table 11.1 presents an implementation of item pre-equating that is similar to common item equating to an item pool except that the new items, *New*, do not contribute to the equating and scoring of X (Kolen & Brennan, 2014). That is, X is assembled from items from a calibrated item pool rather than one specific form and is pre-equated and scored using the IRT parameter estimates from these calibrated items prior to the administration of X . Then, in a subsequent step after the administration of X , scaled IRT parameter estimates are obtained for the *New* items using X and these are added to the item pool for assembling and pre-equating additional forms. Parameter estimates for the item pool and the pre-equating results for other forms assembled from the item pool should all apply to population $P (= T)$. A more complex item pre-equating design involving multiple tests and populations is presented by Petersen et al. (1989), and is a basis of adaptive testing and linking (see the *Linking Adaptive Tests* section).

Pre-equating designs are relatively complex and require stronger assumptions than the other designs described, namely, that the test, section, and item statistics estimated in a pre-equated administration are accurate when used to produce equating and other linking functions. Inaccuracies can result from population differences or from context and order effects where the pre-equating administration involves administering the items and subsections of X in different orders and contexts than those from the actual administration of the intact form X (Davey & Lee, 2011). In addition, for implementations where a typical IRT model is used that assumes test-taker ability is unidimensional, pre-equating could introduce biases when pre-equated IRT statistics based on unidimensionality assumptions are used to approximate a linking for the full multidimensional test (Kolen & Brennan, 2014). Estimates of X from pre-equating could be biased reflections of the actual X , resulting in inaccurate equating results that might be corrected with “postequating” using actual administration data and one of the traditional designs in Table 11.1.

Methodological Implementations of Equating

This section provides a summary of research, discussions, and studies on the methodological issues in equating. The methodological issues covered here pertain to material covered in the following sections: *Data Use and Smoothing*, *Continuization*, *Choices Among Data Collection Designs*, *Approaches to NEAT Equating*, *True Scores*, *IRT*, and *Accuracy Evaluations*. Many of these procedures can and should be implemented using open-source computational programs that provide critical supports for transparency

and for the independent replication of results (Brennan, 2004; Brennan et al., 2009; Gonzalez & Wiberg, 2017). In addition to the methodology covered in this section, important aspects of equating implementations include checks that test that development and administration practices are standardized and similar, that quality control procedures are implemented to ensure administration conditions are followed, and that tests, items, and answer keys are all working and displayed as intended. The implications of altered questions and answer keys are covered by Kolen and Brennan (2014, pp. 331–336).

Data Use and Smoothing

Equating procedures are statistical operations performed on sample data. Ideally, the sample data should come from an established data collection design and should be a large and sufficiently representative sample that can produce precise adjustments for the test linking. The question *How large?* depends on the data collection design and can be addressed based on standard errors of equating for the specific design (Kolen & Brennan, 2014). The question *How representative?* might be answered in terms of data or sample selection procedures, such as exclusion procedures for test takers who are not part of the target population (e.g., from nonrepresentative grades) or who are nonnative English speakers or exam repeaters (Dorans et al., 2011).

Once equating data are obtained, there are additional choices about how to use the data to produce the test-linking results. The most direct choice is simply to use the sample data as originally collected to compute test-linking results. A long-standing interest in the test-equating field is the extent to which equipercentile linking results can be improved by smoothing the test data. The goal of smoothing is to reduce statistical sampling error while not substantially inducing bias. Early methods based on hand smoothing and moving averages of frequencies (Angoff, 1971; Flanagan, 1951) have been replaced with methods that are more accurate, formalized, and efficient.

One smoothing method for equipercentile linking is log-linear presmoothing (Holland & Thayer, 2000). This smoothing method is based on fitting log-linear models to the distributions of the test scores to be linked. The model fitting process produces fitted distributions that match a user-specified number of moments of the unsmoothed data. The process smooths out irregularities and provides plausible nonzero probabilities and frequencies at all possible scores. Log-linear models can be applied to the univariate distribution of the scores of a single test, to the bivariate distribution of two tests, and to the complex structure of a bivariate distribution of a test, an internal anchor and its impossible scores (i.e., structural zeros; H. Kim et al., 2017). Log-linear models can also be used to model other complexities that occur with bivariate distributions of weighted composite scores (Moses, 2014b). Because log-linear models have statistical implementations, they have associated model fit indices including chi-square statistics (e.g., likelihood ratio) and information criteria (Akaike information criterion, Bayesian information criterion, etc.) that lend themselves to statistical selection strategies (Moses, 2011; Moses & Holland, 2010). Finally, estimated variance–covariance

matrices for the smoothed distributions are an output of the modeling that can be used to estimate standard errors of equating and other linking functions that explicitly reflect the smoothing results (Moses & Holland, 2008; von Davier et al., 2004a).

Another smoothing method used in equipercentile equating is known as cubic spline postsmoothing (Kolen, 1984). The “post” in postsmoothing indicates that this smoothing is implemented after an equipercentile X -to- Y equating is computed in order to directly smooth the equating function. The postsmoothing produces adjacent and connected score-level cubic functions that are numerically solved to produce the smoothest possible function that reflects the unsmoothed equipercentile function up to a user-specified smoothing constraint. The implementation, special procedures for the highest and lowest scores, and other procedures for approximating symmetry in the postsmoothed results are described by Kolen and Brennan (2014). Studies have demonstrated that cubic spline postsmoothing and alternatives are useful for reducing error in equipercentile equating functions (Cui & Kolen, 2009). Alternative strategies for selecting postsmoothing parameters have been considered (C. Liu & Kolen, 2018).

Both pre- and postsmoothing introduce bias into equating results. The logic behind both smoothing procedures is to reduce random error without substantially introducing bias. There is no definitive answer to the question, Which is better? It is clear, however, that presmoothing introduces bias before subsequent equating steps are performed, whereas postsmoothing does so when conducted after other equating steps are performed. Also, presmoothing can be automated if one adopts certain prespecified equating criteria, but doing so can be risky without examining smoothed plots before proceeding with equating. By contrast, postsmoothing involves somewhat subjective judgments about smoothing degrees.

In equating practice, data screening, presmoothing, and postsmoothing are important tools for improving the accuracy of equipercentile linking and equating functions. Reasonably large sample sizes are another important matter for ensuring accurate equating results (Dorans et al., 2011; Kolen & Brennan, 2014). When equating data are collected using small sample sizes, equipercentile functions must rely on models that make additional assumptions (e.g., pre-equating data collection designs and assumptions that work under some situations but not others; see Livingston & Kim, 2011).

Continuization

Equating functions based on scale alignment produce a set of “in between” scores that reflect plausible estimates of the difficulty differences between test forms, but that are in fact impossible to obtain. For example, if test X was to be equated to Y and X was easier than Y , then converting number-correct points that adjust for the easiness of X might result in X -to- Y scores that are lower but in between the scores that could actually be obtained (e.g., a score of 28 on X might convert to 27.6). The process of treating discrete test scores as if they are continuous is referred to as “continuizing” (von Davier et al., 2004a). The percentile rank function shown in Equation 11 is one approach to continuizing based on the assumption that the in between scores are uniformly

distributed within boundaries for each score (the score $\pm$ half of the score interval). The linear scale-aligning function (Equation 16) is another approach to continuization based on standard deviation units. The continuization process in equating is described in detail by von Davier et al. (2004a), who showed that the traditional continuization approach based on percentile rank functions produces continuous distributions that can be irregular and can imprecisely reflect the original discrete distribution (i.e., overestimate variance by 1/12).

Von Davier et al. (2004a) described and developed an alternative continuization for equating based on Gaussian kernel smoothing (i.e., kernel equating; Holland & Thayer, 1989). Kernel equating is smoother, is more accurate with respect to variance, and approximates equipercentile, linear, and compromises of these functions through a user-specified kernel smoothing parameter. Several alternative continuization approaches have also been considered, including those based on log-linear models (T. Wang, 2011), exponential families (Haberman, 2011), and alternative kernel functions (Lee & von Davier, 2011). When continuization approaches work to produce values in between two adjacent, attainable scores, different approaches usually exhibit small differences from each other. They can even be shown to be iterative versions of Equation 11's percentile rank computation (Moses & Holland, 2008). One exception is when these approaches are used to equate the highest and lowest scores of a test. Equipercentile equating based on percentile rank functions will connect the highest X and Y scores to each other, whereas equatings based on kernel and linear continuizations may produce results that go further beyond the defined score ranges for the tests (Dorans et al., 2011; Kolen & Brennan, 2014; von Davier et al., 2004a). Another possible exception is scores with unequally spaced values because these are not completely suitable for Equation 11's percentile rank calculations that assume scores with ranges of integers. Flexible continuization approaches could be especially appropriate for tests with unequally spaced score ranges.

Choices Among Data Collection Designs

Each design described in the *Data Collection Designs* section and Table 11.1 presents a particular trade-off of advantages and disadvantages in test equating. These designs can be compared with respect to statistical precision, specifically the sample size of test takers needed to achieve a particular standard error of equating. For a given level of statistical precision, the randomly equivalent groups design and equipercentile equatings require more data than a single group design and linear equating (Kolen & Brennan, 2004). Designs can also be compared with respect to the complications and security concerns they introduce in test administrations (more complexity and security concerns for randomly equivalent groups designs than NEAT designs) and statistical assumptions (minimal assumptions for randomly equivalent groups, more stringent assumptions for NEAT and pre-equating designs; Kolen & Brennan, 2014). Based on the strengths of each design, a recommended ideal design is one in which large samples of test-taker data are collected in a securely implemented randomly

equivalent groups design with an external anchor test that is administered after the tests (Dorans et al., 2011; Holland & Dorans, 2006). This design exploits the simplicity of the randomly equivalent groups design, but also allows for increased statistical precision through use of the anchor test as a statistical covariate. Administering the anchor externally is advantageous in that only the anchor could be affected by context or order effects, and the anchor would not need to be used if these effects were found. Pre-equating designs are usually not recommended based on their complexity and strong assumptions about IRT parameter estimates holding across multiple contexts (Davey & Lee, 2011). However, pre-equating designs could be considered for situations where a test linking must be estimated before that test is actually administered (Kolen & Brennan, 2014).

Approaches to NEAT Equating

Equating texts and studies have given extended attention to the complexities of equating using the NEAT design, considering aspects such as options and designs of the anchor test and computations based on assumptions made about the relationship of the test and anchor scores. One choice is whether the anchor will be internal or external to the tests being equated (Dorans et al., 2011; Holland & Dorans, 2006). External anchors are separately timed, and scores on them do not count toward the X and Y scores.

External anchors or sections have considerable flexibility and multiple uses, such as equating, pretesting, or the tryout of new item types. Potential drawbacks of external anchors are that they must be sufficiently disguised so that test takers do not respond differently to the anchors than they respond to the tests. In particular, test takers should not be able to determine which set of items (or test) is an external anchor, because the goal is that test takers try as hard on the external anchor as they do on X or Y . If that is not true, then scores on the anchor test will almost certainly lead to biased equating results. Recommended practices for equating with external anchors are to use data-screening procedures to identify and exclude test takers with anchor performance that is inconsistent with test performance.

Internal anchors are administered and scored within X and Y , which usually results in higher (anchor, test) correlations than external anchors, but also increases the risk of context and order effects. Recommended practices for equating with internal anchors are to administer the anchor items in similar positions on X and Y and to evaluate and possibly screen the anchor items for differential performance (i.e., differential item functioning).

The predominant recommendation is that external and internal anchors should be designed to be representative of the test, with similar average difficulties (though not necessarily equal difficulty spread; Sinharay & Holland, 2006a, 2006b). The anchor(s) should also be relatively long and reliable (Moses & Kim, 2007), where the long-standing recommendation is for at least 20 items or 20% of the items on the test (Angoff, 1971). Also, they should be administered and screened to reduce atypical test-taker performance, context, and order effects. Following these guidelines helps ensure that

the anchors will be as highly correlated as possible with the tests being equated, which supports accurate NEAT equatings.

Once the NEAT data are obtained for a sample of population P that takes X and A and a sample of population Q that takes Y and A (Table 11.1), different options are available for using the anchor data to equate X and Y in T . Two major approaches involve chaining scale-aligning functions through the anchor and computing scale-aligning functions from test distributions estimated for a single hypothetical group on both X and Y .

The chained approach involves computing a scale alignment function from X to A in the P data and another scale alignment function from A to Y in the Q data and chaining them together. The assumption is that both scale alignment functions are population invariant (i.e., apply to other populations, and specifically to target population T ; Holland & Dorans, 2006). Expressed in equipercentile functions (Equation 15) with percentile rank functions for A in P and Q , $H_P(a)$ and $H_Q(a)$, and linear functions (Equation 16), the chained equipercentile and chained linear scale alignment functions are

$$e_{XT}(x) = G_Q^{-1}(H_Q\{H_P^{-1}[F_P(x)]\}) \quad (21)$$

and

$$l_{YT}(x) = l_{YQ}[l_{AP}(x)], \quad (22)$$

where $l_{YQ}[\]$ denotes the A -to- Y linear function in the Q data.

The second approach (scale aligning) is referred to as frequency estimation or poststratification and involves projecting (see the *Distinguishing Equating From Other Forms of Linking* section) the X and Y distributions conditional on A for T . The projection is based on assumptions that the conditional $X|A$ distribution observed in population P applies to population Q and that the conditional $Y|A$ distribution observed in population Q applies to population P . The resulting X and Y distributions are estimated for a synthetic, combined group of P and Q test takers that comprise the target population $T = w_P P + w_Q Q$ ($w_P + w_Q = 1$, and w_P and w_Q range from 0 to 1),

$$Pr_T(x) = \sum_a Pr_P(x|a)[Pr_T(a)] = \sum_a Pr_P(x|a)[w_P Pr_P(a) + w_Q Pr_Q(a)] \quad (23)$$

and

$$Pr_T(y) = \sum_a Pr_Q(y|a)[Pr_T(a)] = \sum_a Pr_Q(y|a)[w_P Pr_P(a) + w_Q Pr_Q(a)]. \quad (24)$$

The X and Y distributions produced with Equations 23 and 24 are estimated for the same target population and can be used in percentile rank and scale-aligning functions like Equations 11 and 15. Different linear scale-aligning functions based on poststratification are possible, such as those that use the means and standard deviations from the distributions obtained from Equations 23 and 24 in Equation 16 (Braun & Holland, 1982). Another method attributed to Tucker (Angoff, 1971) makes assumptions that the conditional means of the tests are linear and invariant given anchor scores and that the conditional variances of the tests are constant and invariant given anchor

scores, producing estimated means and variances of X and Y on T from anchor-to-test regression equations like Equation 17 (Gulliksen, 1950; Lord, 1950).

The chained and poststratification approaches to equating tests with the NEAT design have been compared in several research studies. Von Davier et al. (2004b) showed that these methods produce identical results when the (anchor, test) correlations are perfect or when the anchor distributions are identical in the P and Q samples. These results underscore the view that in ideal equating situations, such as those where (anchor, test) correlations are high and where administration groups are similar, different equating methods can produce very similar results (Dorans et al., 2011). Studies have also considered situations where the chained and poststratification methods give different results, where a choice of equating method is more consequential for reported scores. Summaries of these studies indicate that the poststratification approaches have smaller standard errors than the chained approaches and less overall equating error when group differences are small, whereas the chained approaches are less biased and have less overall equating error when group differences are not small (Dorans & Puhan, 2017; Kolen & Brennan, 2014; Kolen & Lee, 2011, 2012, 2014, 2016, 2018).

True Scores

The equating of true scores is a long-standing theoretical interest. This interest involves equating procedures based on theory and measurement models, especially classical test theory (see the *Equating Requirements* section, Equation 19; Angoff, 1971; Levine, 1955). For the NEAT design, the test forms and anchors are usually assumed to follow a classical congeneric model, and versions of Equation 19 are used to compute linear τ_X -to- τ_A and τ_A -to- τ_Y functions that are chained together for a Levine equating of τ_X and τ_Y (Hanson, 1991; Kolen & Brennan, 2014; Levine, 1955),

$$l_{\tau_Y}(\tau_X) = l_{\tau_Y}[l_{\tau_A}(\tau_X)], \quad (25)$$

where $l_{\tau_Y}[\]$ denotes the τ_A -to- τ_Y linear function. The relationships of the true test and anchor scores are assumed to be population invariant (Holland & Dorans, 2006).

Equating functions for true scores address theoretical interests while introducing conceptual difficulties for practice. The main difficulty is that true scores are unobserved and unavailable in practice (Lord, 1980; Lord & Wingersky, 1984) so that true score equating functions like Equation 25 typically use observed scores, $l_{\tau_Y}(x)$. Although the interchangeability claims that true score equating functions support apply to true scores, the equity discussions described in the *Equating Requirements* section show that specific types of equity are achieved in observed score applications when test forms and anchors are classically congeneric. Still, such applications are inconsistent, lack compelling reasons (Kolen & Brennan, 2014), and do not reflect the measurement error in X that is unaccounted for when equating is attempted with true score conversions (especially beyond the conditional means addressed in first-order equity). From the *Equating Requirements* section, the strongest connections in theory and practice are achieved

when the test forms and anchors are constructed and administered in the same way (reflecting congeneric theories and first-order equity) and the test forms are equally and highly reliable (such that observed scores approximate true scores and $l_{TY}(x)$ functions reflect second-order equity).

Another version of Levine linear equating is based on the same assumptions for X , Y , and A and their true scores as used in Equation 25, but equates the observed X and Y scores in a hypothetical target population (defined with respect to true anchor scores) using frequency estimation and poststratification assumptions about X and Y scores given the true anchor scores (Holland & Dorans, 2006; Kolen & Brennan, 2014; Levine, 1955). Levine equating has been shown to perform well in some situations where administration group differences are large (Kolen & Brennan, 2014; Mroch et al., 2009), though these results depend on the accuracy of the Levine assumptions.

Equipercntile scale-aligning functions based on true scores have been developed by Chen et al. (2011) and T. Wang and Brennan (2008). T. Wang and Brennan's (2008) proposal is to modify the frequency estimation equipercntile procedure in Equations 23 and 24 so that these equations are based on a projection from estimated true anchor scores (i.e., $Pr_p(x|\hat{\tau}_A)$). Chen et al.'s (2011) approach generalizes chained and post-stratification equipercntile approaches developed in the kernel equating framework so that these are based on true scores.

IRT

IRT can also be used to produce equating functions of observed or true test scores or θ estimates. IRT applications require that the IRT assumptions hold and that parameter estimates for X and Y items are on a common scale. When these item parameters for the X and Y tests are estimated with a single population design such as the randomly equivalent groups data collection design (Table 11.1) using the same specification of the θ scale, the parameter estimates can be interpreted as being on the same scale. When tests are administered to nonequivalent groups as in the NEAT design, procedures are needed to transform the parameter estimates to a common scale, either through special linear transformations based on the two sets of IRT parameter estimates for the anchor items or through concurrent calibrations of the X and Y test items and anchor items (Kolen & Brennan, 2014).

For IRT observed score equating, an IRT-based recursive algorithm can be used to estimate probability distributions for X and Y given θ , $Pr(x|\theta)$, and $Pr(y|\theta)$ (Lord & Wingersky, 1984). Once obtained, these conditional distributions are averaged over a target distribution of θ ; specifically,

$$Pr_T(x) = \int Pr(x|\theta) Pr_T(\theta) d\theta \quad (26)$$

and

$$Pr_T(y) = \int Pr(y|\theta) Pr_T(\theta) d\theta. \quad (27)$$

Equations 26 and 27 can be described as IRT-based presmoothed distributions (Holland & Dorans, 2006), which can be used in equipercentile equating. IRT-based true score equatings can also be produced (see the *Distinguishing Equating From Other Forms of Linking* section). Finally, if a testing program's scale scores are produced using θ as the test performance measure (see the *Scaling Methods for Primary Scale Scores* section), equating procedures may be bypassed altogether and test takers' $\hat{\theta}$ s may be estimated by applying IRT ability estimation procedures to their item-level performance (see the *IRT Ability Estimates* section) and scaling these estimates to the reporting scale, $sc(\hat{\theta})$.

Accuracy Evaluations

Once estimated, test equating functions should be evaluated for accuracy. Most of the accuracy evaluations reviewed in this section are comparative, where an equating function based on one method and associated assumptions is evaluated in comparison to one or more other equating functions based on different methods and assumptions. For example, an equipercentile equating function might be of interest and evaluated in terms of how different it is from a simpler equating approach, such as a linear equating, or the use of raw scores to consider whether equating is needed at all. In subpopulation invariance evaluations, equating functions obtained from subgroup data are typically compared to a function based on the total group, usually for purposes of evaluating subgroup dependencies. These comparisons require the estimation of two or more plausible equating functions and evaluating whether their differences are "large," based on material covered in the following sections: *Differences That Matter*, *Standard Errors*, *Subpopulation Invariance Evaluations*, *Correlations and Prediction Error*, and/or *Evaluations and Recommendations for Drift in Equating Chains*.

Differences That Matter

Evaluations of the magnitude of equated score differences are based on whether those differences are big enough to alter test takers' reported scores. Criteria for these differences, known as "differences that matter" (Dorans & Feigenbaum, 1994), are traditionally defined based on the differences that are considered so large that they are not eliminated when reported scores are rounded, such as half of the integer of the equated scores (0.5 points for equatings based on raw summed scores) or half of the unit of the scale score interval. These values can provide benchmarks for evaluating the magnitude of equated score differences based on two equating functions, but because very similar equated scores can round to different values, the differences-that-matter criteria should not be treated dogmatically (Kolen & Brennan, 2014). Clarifications were provided by J. Liu and Dorans (2012), who proposed a procedure for evaluating whether test takers who take a test under altered conditions should be scored based on a conversion specifically produced to reflect those altered conditions. Liu and Dorans's procedure distinguished four categories of equated score differences, comprising a 2

$\times 2$ table of rounded reported scores (are they the same or different for two possible conversions?) and the magnitude of the differences in the unrounded scores (are they trivial or nontrivial?). Based on these distinctions, Liu and Dorans recommended that a separate equating for scores based on the altered test conditions should be considered only when more than half of the test takers whose reported scores are altered (when the separate conversion is applied) also have nontrivial differences in their unrounded scores.

Standard Errors

Standard errors of equating can be useful for evaluating the extent to which potential equating functions differ by more than can be attributed to test-taker sampling errors. Standard errors of linear equating functions have a history that can be traced to Lord (1950), with updated discussions reflective of equipercentile and kernel functions, different data collection designs, distributional assumptions, and log-linear presmoothing models (Angoff, 1971; Braun & Holland, 1982; Jarjoura & Kolen, 1985; Liou & Cheng, 1995; Lord, 1982; Moses & Holland, 2008; Moses & Zhang, 2011; Ogasawara, 2001, 2003; von Davier et al., 2004a). The standard errors of equating are defined relative to the standard deviation of the distribution of equated scores produced from random samples of test takers from their respective populations,

$$SEE_{YT}(x) = \sigma[\hat{e}_{YT}(x)]. \quad (28)$$

Standard errors of equating are usually based on assumptions that the tests, their items, test-taker sample size(s), and other decisions made in the equating (e.g., smoothing) are fixed and not varied across replications of the test-taker sampling. Most of the cited sources present standard errors that are asymptotically derived, though standard errors can also be estimated through simulations and resampling approaches (Kolen & Brennan, 2014).

Standard errors for a single equating function might be used in evaluations of equated versus raw score differences, essentially questioning whether test score equating affects test scores more than can be attributed to statistical error from the test-taker sample(s). One application is estimating the test-taker sample size(s) needed to obtain a given level of equating precision (Kolen & Brennan, 2014, pp. 273–276). Another application is the situation where a testing program might offer different editions of its test in which the items, item sets, and/or test sections appear in different orders. The extent to which order effects are observed in test scores can be evaluated through equating the scores of the alternate editions to each other and evaluating how different the equated and nonequated test scores are. To the extent that the differences are “large,” the equating function used to evaluate the item order effects might also be used to adjust scores for the order effects. Standard errors for an equating function provide a basis for evaluating whether equated and nonequated scores differ more than can be attributed to test-taker sampling. These ideas were used to evaluate order effects in the SAT (Dorans & Lawrence, 1990) and in Advanced Placement (Moses et al., 2007).

For questions about differences in equated scores from two plausible equating functions, each of which reflects sampling error, standard errors of equating differences (SEEDs) have been developed (Moses & Holland, 2008; Moses & Zhang, 2011; Moses et al., 2010; von Davier et al., 2004a),

$$SEED_{Y,T,1,2}(x) = \sigma[\hat{e}_{Y,T,1}(x) - \hat{e}_{Y,T,2}(x)]. \quad (29)$$

SEEDs can be used to evaluate differences between two equating functions, addressing several possible questions. Equated score differences of interest include curvilinear and linear kernel equating functions (von Davier et al., 2004a), traditional equipercentile and linear equating functions (Moses & Zhang, 2011), poststratification equating functions obtained from different target populations (von Davier et al., 2004a), counterbalanced design equating functions based on different weights of the samples taking the tests in different orders (von Davier et al., 2004a), poststratification equipercentile and chained equipercentile equating functions (von Davier et al., 2004a; Moses & Holland, 2008), and poststratification equating functions based on one or two anchor scores (Moses et al., 2010).

Subpopulation Invariance Evaluations

The subpopulation invariance requirement of equating (see the *Equating Requirements* section; Equation 20) as described by Dorans and Holland (2000) prompted several empirical evaluations of the sensitivity of linking results with respect to subpopulations (Dorans, 2004a; von Davier & Liu, 2007). Recent studies indicate renewed interest in a long-standing question in test equating (Kolen, 2004), one that can be traced to Flanagan's (1951) statements that group dependencies in equating are to be expected (p. 748) and to Angoff's (1971) statements that equating functions should be independent of the individuals used to compute them (p. 563). The inclusion of subpopulation invariance as a requirement for test equating corresponds to research and expectations for subpopulation invariance in equating. Empirical checks are important for determining whether test linkings exhibit desired equating properties and for evaluating intended equating results for tests that undergo transitions, such as in their specifications and possibly in their constructs (J. Liu & Dorans, 2013).

Consider the subpopulation invariance measure in Equation 20, for which Dorans and Holland (2000) developed two measures to quantify the lack of invariance in intended equating functions for an exhaustive set of mutually exclusive subpopulations:

$$RMSD(x) = \frac{\sqrt{\sum_{T_g} w_g [e_{Y,T_g}(x) - e_{Y,T}(x)]^2}}{\sigma_{Y,T}} \quad (30)$$

and

$$REMSD = \frac{\sqrt{\sum_{T_g} w_g E\{[e_{Y,T_g}(x) - e_{Y,T}(x)]^2\}}}{\sigma_{Y,T}}, \quad (31)$$

where the $\mathbf{E}$ denotes an expected value and where w_g is the proportional sample size for T_g in T . The basis of the $RMSD(x)$ and $REMSD$ measures is quantifying the lack of subpopulation invariance in test equating functions of T_g from that of T , at the individual score level and also at an overall level. In their discussion of the $RMSD(x)$ and $REMSD$ measures, Kolen and Brennan (2014) suggested that these measures could mask other issues, such as the extent to which the test linking results for two subpopulations differ from each other. They proposed several additional measures that could directly address this question based on examining

$$e_{Y_{T1}}(x) - e_{Y_{T2}}(x). \quad (32)$$

Additional measures of invariance can also be computed for specific subgroups of interest (J. Liu & Dorans, 2013),

$$RES_{D_{T_g}} = \sqrt{\sum_x Pr_{T_g}(x) [e_{Y_{T_g}}(x) - e_{Y_T}(x)]^2}, \quad (33)$$

where $Pr_{T_g}(x)$ is the relative frequency at $X = x$ for subgroup T_g .

Subpopulation invariance investigations were conducted using Equations 30–33 and other measures in Dorans (2004a), von Davier and Liu (2007), Kolen and Brennan (2014), J. Liu et al. (2010), and Yin et al. (2004). Standard errors for these measures were developed and studied by Moses (2008) and Rijmen et al. (2009).

One suggestion for practice is to apply subpopulation invariance concepts to conduct population invariance evaluations that address invariance for populations that are intended applications of an equating function but not necessarily represented in the actual equating study. Experience suggests that some of the most serious problems with equating and linking results involve estimating linking functions in data from one group (e.g., special studies, small subgroups of an administration used for randomly equivalent groups designs, pre-equated conversions) and applying these linking functions to other groups, testing conditions, and administration data. These experiences led to recommendations that equatings and linkings conducted in special studies with nonrepresentative scoring, administration conditions, and test-taker data should be regarded as limited in the score interpretations they suggest and subject to additional research to support their use and interpretations with the general testing population(s) (W. Lee & Brennan, 2021; Moses, 2022). This additional research can include invariance investigations for different groups as well as special data reviews that ensure that the test equating is working reasonably well for particular users and administration groups. These checks could be as simple as reviewing the scale score distribution(s) from applying an estimated equating function to the administration group and comparing it to historical distributions for that group.

Correlations and Prediction Error

Equating and other linking functions can be produced to support multiple purposes and interpretations, such as to align test scales based on score distributions and to

provide predictions of test takers' scores. The scale-aligning computations used to support interchangeability across alternate forms in equating, or comparability across distinct tests in concordance studies (see the *Concordances* section), might have other intended uses as predictions of test takers' performance on those alternate forms or tests. To the extent that test correlations and reliabilities are high (see the *Distinguishing Equating From Other Forms of Linking* section), scale-aligning functions are more likely to support both aligned scales and accurate predictions. As test reliabilities and correlations decrease, the differences between scale-aligning functions and prediction functions increase, such that the predictions from scale-aligning results are increasingly biased. If the scales of independent random numbers were aligned, the results would have no predictive utility because they would not reflect the reliabilities and correlations of the "scores" (Dorans, 2004b).

When describing these issues, Dorans (1999) used a coefficient of alienation based on the prediction error of regression functions,

$$\frac{\sigma_{YT} - \sigma_{YT} \sqrt{1 - \rho_{XY,T}^2}}{\sigma_{YT}} = 1 - \sqrt{1 - \rho_{XY,T}^2}. \quad (34)$$

In words, Equation 34 indicates the proportion that uncertainty is reduced from predicting Y from using a linear X -to- Y regression function compared to using the mean of Y . Dorans and colleague (Dorans, 1999; Dorans & Walker, 2007) argued that concordances should only be produced when scores for tests X and Y are correlated at least .866, in which case Equation 34 indicates a 50% reduction in prediction uncertainty in standard deviation units using an X -to- Y regression. For tests correlated less than .866, scale-aligning methods result in biased predictions and should be replaced with prediction or projection methods (Dorans, 1999, 2004b). Although the correlation in Equation 34 suggests the need for a data collection design where test takers take both X and Y , Equation 34 can be used to evaluate equating results from other designs where X and Y are alternate forms assumed to be parallel. For the equating of alternate forms, the XY correlation can be estimated from the test reliabilities (Gulliksen, 1950; Lord & Novick, 1968).

These discussions were updated by Moses (2014a), who provided analogues to Equation 34 based on the prediction error from a linear scale-aligning function (Equation 16),

$$\frac{\sigma_{YT} - \sigma_{YT} \sqrt{2[1 - \rho_{XY,T}]}}{\sigma_{YT}} = 1 - \sqrt{2[1 - \rho_{XY,T}]}. \quad (35)$$

Equation 35 indicates that equatings, concordances, and other linkings produced from linear scale-aligning functions reduce prediction uncertainty by 50% when X and Y are correlated at least .875. Results were provided by Moses (2014a) to compare linear regression and scale alignment functions in terms of prediction error, showing that these

functions are similar for high XY correlations but that linear regressions are increasingly favored when correlations between X and Y are low. Additional insight can be obtained by decomposing prediction error variances for regression and scale alignment functions into their proportional contributions of true score and error variances (Moses, 2014a) and by using other related measures of proportional reductions in mean-square error (Dorans, 2022).

EVALUATIONS AND RECOMMENDATIONS FOR DRIFT IN EQUATING CHAINS Established testing programs must produce several editions of a test that are developed, administered, equated, and used to score test takers. This history means that testing programs can have several previously administered and equated forms potentially available for the equating of additional test forms. Such programs would also have historical expectations about the reasonableness of scale score distributions for given administration groups. An increasing number of test equatings back to the scale is expected to cause drift due to the accumulation of random error (Kolen & Brennan, 2014). Large numbers of previous equatings also provide opportunities for monitoring historical equating results and for evaluating the consistency of current equating and scale score results.

One way that testing programs can improve test equating accuracy is to conduct their equatings to link their tests back to two or more previously equated forms (Holland & Dorans, 2006; Kolen & Brennan, 2014). From 1994 to 2016, the SAT conducted external anchor equatings to link new forms back to four old forms. Multiple links to past forms provide ways to detect aberrant equating results and scale score conversions. Multiple links also make the final equating and scale score conversions less reliant on any one previously developed equating that may be problematic.

Established testing programs can also monitor their current and historical equating results for evidence of scale score drift. Modu and Stern (1975) monitored SAT scales from 1963 to 1973, finding evidence that the verbal and mathematics sections had drifted. More recently, Haberman and Dorans (2011) delineated several contributors to scale score inconsistency, including anchors for the NEAT design, sampling errors (random and nonrandom), accumulated random error, and model misfit. Some practices were noted to exacerbate scale drift, such as continuous testing where more new forms are administered and equated with smaller groups of test takers (increasing standard errors and the accumulation of standard errors in chains of equatings). When equatings and raw-to-scale conversions are available for several administered forms, additional methods for evaluating drift have been described, such as harmonic regression and time series analyses for evaluating seasonality effects in scale score conversions (Y. Lee & Haberman, 2013) and quality control charts and time series methods to support continuous monitoring, adjustment of variations, identification of abrupt shifts, and the assessment of autocorrelation. More recent evaluations of scale stability in the SAT have also been conducted, such as studies of the extent to which drift in equating chains is affected by different degrees of postsmoothing (S. Y. Kim et al., 2020)

and the drift due to the accumulation of random and nonrandom errors observed from readministering and re-equating an old form and comparing this to more recently equated forms (Guo et al., 2012). Another study evaluated the extent to which scale score drift and variability in the SAT were controlled by equating back to two or three old forms rather than one (J. Liu et al., 2014). These studies provide examples of how equating drift might be assessed and potentially reduced.

LINKING

In the *Equating* section, equating was described as the strongest type of linking, subject to the strictest requirements, used to adjust the scores of alternate forms of a single test for unintended difficulty differences for the purpose of establishing interchangeable scores. The *Linking* section describes other types of linkings for the scores of tests that are distinct and not expected to meet all of the equating requirements of equal constructs, same test specifications, equal and high reliabilities, same administration conditions, etc. Linking efforts might be undertaken to promote particular interpretations, such as appropriate comparisons or predictions from distinct tests. These goals are not as ambitious as the interchangeability goal of equating. Although the comparability established in linking might be general in intention (e.g., score comparisons across different tests and testing contexts), the results of linking distinct tests are nevertheless weaker (i.e., less precise) than equated results, with interpretations that must usually be qualified and limited in some way. For example, the score comparability achieved through scale aligning methods is considered to be “assured only for that specific group taking the tests under specific conditions (Angoff, 1971)” (Pommerich, 2016, p. 117). Predictions and regressions are also noted to be group specific (Linn, 1993; Mislevy, 1992).

This section summarizes types of linking other than equating, their uses, and their limitations. The linking types covered in earlier linking frameworks include those in the *Concordances*, *Vertical Scaling*, *Battery Scaling and Composites*, and *Predicting and Projecting* sections, all of which are summarized in Table 11.3 based on a tabled version of Holland and Dorans’s (2006) linking framework. Some linking types described in previous discussions are omitted because of ambiguities and updates in terminology (*calibration* was part of earlier linking frameworks, but is avoided here because it sometimes refers to vertical scaling and other times to fitting IRT models; Holland & Dorans, 2006). Approaches described in the *More Recent Linking Types* section are also covered, including linking for exams administered in different testing modes, linking subjectively scored tests, local linking, linking adaptive tests, linking exams administered to different test takers using weak anchors, and linking state tests to NAEP (Table 11.4).

Concordances

Concordances are scale-aligning linkings for tests built from similar but not identical test specifications, where the tests have similar uses, lengths and reliabilities. One

Table 11.3 Summary of Nonequating Types of Linking

Type of Linking	What Is Dissimilar?	What Is Similar?	What Is Equal?	Likely Data Collection Design	Computation	Interpretations
Concordance		Constructs; reliability; difficulty	Population	Single group; approximated counterbalanced	Scale aligning	Comparable scales for the overall concordance group, but not necessarily to subgroups
Vertical scaling	Difficulty; populations	Constructs; reliability		NEAT with common items or scaling test, or random groups	Scale aligning	Comparable scales that depict growth based on the methodological choices for how the vertical scales were established
Battery scaling	Constructs		Population	Single group or randomly equivalent groups		Comparable scales that indicate test takers' relative performance across the test battery
Composite scales			Population	Single group taking the individual tests	The composite is a sum of the scale scores of the individual tests, which are maintained through scale-aligning equatings	Composite scales are indirectly maintained to the extent that the scales of the individual tests are maintained and the intercorrelations of the individual tests do not change
Predicting and projecting			Population	Single group; counterbalanced	Predicting and projecting	Best prediction for a prediction group

Table 11.4 Summary of Recent Linking Examples

Type of Linking	What Is Dissimilar?	Likely Data Collection Design	Computation	Interpretations
Linking across conditions of measurement	Conditions of measurement	Various	Scale aligning	Comparable scales, with caveats about limitations when testing mode differences are large
Linking subjectively scored tests	Scoring that drifts	Various	Scale aligning	Comparable scales, depending on high reliability and procedures to account for trends and rater drift across administrations

(continued)

Table 11.4 (continued)

Type of Linking	What Is Dissimilar?	Likely Data Collection Design	Computation	Interpretations
Local, ability-specific test linking	A form's scale alignments at different abilities	Various	Conditional scale aligning	θ -Conditional comparable scales
Linking adaptive tests	Forms and form difficulties at different abilities	Pre-equating	IRT ability estimation	Comparable scales at specific levels of adaption, influenced by the item pools, and the accuracy of the IRT models, scores and adaptations
Linking tests using weak anchors	Test-taker groups; collateral information vs. the tests	NEAT design without a suitable anchor	Scale aligning on projected distributions	Comparable scales for hypothetical groups defined by the weak anchor(s)
Linking state tests to NAEP	State tests and NAEP	Single group or counter-balanced	Scale aligning or projecting	Comparable scales or projected results are more supported within states than across them

Note. IRT = item response theory; NAEP = National Assessment of Educational Progress; NEAT = nonequivalent groups with anchor testing.

example is the concordance of the SAT and ACT reporting scales. Other concordance discussions and examples involve the ACT and ITED tests (Yin et al., 2004), and, arguably, those described in the *Linking State Tests to NAEP* section. Concordances are intended to align the tests' scales using scale-aligning functions and support additional desires to use the resulting concordance tables as surrogates for test takers' scores on the test they did not take.

Early arguments about concordances for the ACT and SAT were summarized by Pommerich (2007), including concordance table proposals for avoiding excessive testing in schools and arguments from Angoff (1962) and Lindquist (1964, February) against concordance tables because of problems in these proposals and possible misuses of their results. Since these early arguments, several ACT/SAT concordances have been produced (College Board/ACT, 2018; Dorans, 1999; Dorans et al., 1997; J. Liu et al., 2010; Marco & Abdel-Fattah, 1991). These ACT/SAT concordance studies are based on obtaining data from test takers taking the ACT and the SAT at least once, in either order, selecting the ACT and SAT scores to use for the test takers who take either test more than once and screening the test-taker data to avoid large time differences between ACT and SAT testings. Linking results are produced from the screened data as equipercentile conversions for some combinations of ACT and SAT tests (e.g., the ACT Summed score and the SAT Total scores in 2018, the ACT Summed score and the SAT Verbal + Math

score in 2010). Decisions about which ACT and SAT tests to concord are usually based, in part, on the size of their correlations (see the *Correlations and Prediction Error* section; Dorans, 1999). Content similarities and political issues often inform these decisions as well. The usual data collection design is a single group design that ignores testing order. In some ACT/SAT concordance studies, testing order is accounted for by weighting data for each order similar to a counterbalanced design, but for nonequivalent groups (College Board/ACT, 2018; Marco & Abdel-Fattah, 1991). Other statistical weighting procedures have also been used to improve the representativeness of the concordance data and resulting concordances for the test-taker populations of each test, who do not necessarily take both tests (College Board/ACT, 2018).

Because of differences in the specifications of the tests in concordance studies, concordance results are not expected to provide interchangeable scores. In fact, concordance tables are expected to be group dependent. For example, the concordance of the ACT Summed score and the SAT Verbal + Math score differs for females and males and across different race/ethnicity groups. These differences can be attributed in large part to the fact that math content contributed 25% to the ACT Summed score and 50% to the SAT Verbal + Math score (Dorans, 2020). In addition, concordances produced from very specialized conditions of measurement, scoring, or with restricted samples or in special studies may not generalize to entire testing populations and may warrant additional studies to update results (Dorans & Moses, 2023; W. Lee & Brennan, 2021; Moses, 2022).

Vertical Scaling

Vertical scaling is used to establish “developmental” scales for reporting performance on versions of a test that are appropriate for specific ages or grade levels represented in schooling, and usually K–12 testing (Kolen, 2006; Kolen & Brennan, 2014). In Holland and Dorans’s (2006) linking framework, vertical scaling involves the linking of tests that measure similar constructs at similar reliabilities, but that differ in difficulty and in the test-taker populations. Thurstone (1925, 1938) proposed vertical scaling methods for item difficulties and then modified them for summed scores (later described as a method of absolute scaling for age and grade-based scales; Flanagan, 1951; Gulliksen, 1950). Rasch IRT models were considered for vertical scales in the 1970s and 1980s (Briggs & Weeks, 2009). More recent vertical scales include CTB/McGraw–Hill’s Terra Nova (CTB/McGraw–Hill, 2001), the Iowa Test of Basic Skills (Hoover et al., 2003; Kolen 2006; Petersen et al., 1989), the ACT scales (Brennan, 1989), tests of English acquisition and English as a second language (ETS, 2005; J. Wang & Smith, 2003), and the 2016 redesigned SAT Suite (Y. K. Kim et al., 2016). Typically, a separate vertical scale is established for each test in an overall testing battery, serving as potential inputs to battery scaling (see the *Scales for Test Batteries and Composites* and *Battery Scaling and Composites* sections). The resulting developmental scales and scores are used to provide a means by which students’ growth is measured, observed, and used to plan instruction and instructional support across schooling. Because the tests being linked in the

vertical scale differ in difficulty, length (usually), timing, content, and other aspects, vertical scaling results do not produce interchangeable scores across levels (Kolen & Brennan, 2014).

Several issues affect the production and results of vertical scales (Kolen, 2006), including definitions of growth, grade-based testing content, data collection designs, linking methodologies, and implementation choices in an IRT model, scoring, and estimation.

How Growth Is Defined

Vertical scalings can reflect different definitions of growth, such as *grade to grade*, where growth is defined over the test content appropriate to particular grades (usually two adjacent grades), and *domain*, referring to growth over all content in the domain (Kolen & Brennan, 2014).

Grade-Based Testing Content

For subject matter areas closely tied to a school curriculum, students tend to exhibit different amounts of growth depending on the content areas on which students are tested. For example, if division by whole numbers is taught in Grades 3 and 4, then growth in the third and fourth graders is expected to be greater in this area than growth in fifth and sixth graders (Kolen & Brennan, 2014). Different vertical scaling results might result from conducting the scaling on either grade (Holland, 2007, p. 18). Growth rates would be different across vertically scaled tests measuring domains that are differentially associated with school curriculum. Vertical scaling results can also be affected by how well the difficulty levels are represented in tests for earlier and later grades (Kolen, 2006, p. 178).

Data Collection Designs

Three designs are usually contrasted for vertical scaling (Kolen & Brennan, 2014). For vertical scales established with the NEAT design, adjacent grade-based tests are taken by students from those corresponding grades, and the blocks of items representing overlapping content for those two grades serve as the anchor test. The randomly equivalent groups design can also be used, where test takers in each grade are randomly assigned to take either a grade-specific test or a test designed for an adjacent grade. The NEAT and equivalent groups designs both involve the chaining of test linkings for tests administered to adjacent grades, and both designs reflect a grade-to-grade definition of growth.

A third data collection design is the scaling test design, which is similar to the NEAT design where an external “scaling” test that is built to reflect content from the entire domain and represent all of the grades of interest is used as the anchor. Typically, the scaling test is administered to students along with the grade-specific tests to be scaled. The scaling test design reflects a domain definition of growth. The scaling test design has been used with the Iowa Test of Basic Skills, ACT, and SAT testing programs.

Linking Methodologies

Vertical scales can make use of methods based on several measures of test performance (Kolen, 2006). For summed scores, Hieronymous scaling involves the linking of summed score medians across tests. Thurstonian scaling is based on linking grade-based summed score distributions after they are transformed into normal distributions. The untransformed raw scores could also be vertically scaled, such as through chained equipercentile procedures (Y. K. Kim et al., 2016). IRT scoring methods can also be used to establish vertical scales.

IRT Implementations

The use of IRT for vertical scaling involves choices of an IRT model, estimation choices, and the IRT-based scores to use (Briggs & Domingue, 2013; Briggs & Weeks, 2009; Kolen, 2006). IRT models such as the Rasch, 2PL, or 3PL models could be used. Test items' IRT parameter estimates could be obtained through calibrations conducted separately for each grade-specific test, items, and test-taker data or concurrently for the tests, items, and test takers from all grades. Vertical scales could be established from different proficiency estimators, including MLE (Equation 4), EAP (Equation 5), or TCC approaches (Equation 6).

The test score linkings resulting from vertical scales have been characterized as a “folding ruler” and as a “ruler that bends” (Yen, 2007, pp. 274–275). What causes vertical scales to fold and bend? As noted previously, the grade-specific tests can show more growth for content that is specifically taught in the curriculum for those grades (perhaps suggestive of grade-to-grade growth). Reviews of vertical scales from the 1980s describe a shrinkage phenomenon where the scores from several types of vertical scales tended to have standard deviations that shrunk from earlier to recent tests and grades (Briggs, 2013; Kolen, 2006; Yen, 2007). Shrinkage was most often shown in vertical scales established with IRT, potentially affecting the ordering of schools based on their gain scores (Briggs & Domingue, 2013). Shrinkage in IRT scales results in a depiction of growth that differs from the growth depicted in vertical scales established with non-IRT approaches, where shrinkage suggests more rapid growth for lower achieving students who catch up to higher achieving students (Briggs, 2013; Hoover, 1984). Although the source of the scale shrinkage in vertical scales from the 1980s was never definitively established, suggested explanations point to early IRT estimation methods (i.e., joint maximum likelihood estimation), multidimensionality within and across the tests, and a failure to establish interval scales. Vertical scales since the 1980s have generally not exhibited shrinkage (Yen, 2007), coinciding with updates in IRT procedures.

Additional influences on vertical scales have been elaborated (Briggs & Weeks, 2009), including the chosen IRT model (scales based on 3PL models result in greater scale score variability than those based on Rasch models), IRT estimation (MLE results in greater score variability than EAP), and calibration (concurrent calibration decreases scale variability vs. separate calibration). IRT limitations can also affect vertical scales, such as strong and potentially unrealistic assumptions that item parameter estimates

fit item response data and are invariant across ages and grades, as well as assumptions that items and test-taker abilities can be modeled as unidimensional. Multidimensional approaches that account for changes in constructs measured across grades can show limitations with unidimensional approaches (Weeks, 2018). For these and other issues described in the cited sources, vertically scaled scores across grade- and age-specific tests can produce inconsistent estimates of test takers' growth, estimates that reflect choices in methodology, test development, and administrations. Altogether, these results indicate that scores from different tests in the vertical scale are not interchangeable.

Battery Scaling and Composites

Battery scaling can be described both as an approach to establishing scales (see the *Scales for Test Batteries and Composites* section) and as an indirect way to link the scores of a battery of tests designed to measure different constructs that are administered to a common population of test takers (Holland & Dorans, 2006). The purpose of establishing the scales in a similar way for all the tests in a battery is so the resulting scales can facilitate interpretations about relative performance and strengths and weaknesses across the battery. Examples include an early proposed battery scale for different scoring methods for handwriting (Kelley, 1914) and for establishing scales for recent versions of the SAT and ACT batteries (Brennan, 1989; Dorans, 2002; Y. K. Kim et al., 2016). It is likely and expected that the linkings of test scores produced in battery scaling do not produce interchangeable test scores. One way that this lack of interchangeability can be observed is in subpopulation dependencies, such as for subgroups of test takers who do relatively better (or worse) on a mathematical measure and also do worse (or better) on a verbal measure.

Composite scores are usually derived from individual tests in a battery, making the comparability of composite score scales indirect and difficult to maintain. To illustrate some challenges with composite score scales, consider a situation where the scales of two tests are established and maintained through equating, and let these two tests for which alternate forms would be developed be represented as Y_1 and Y_2 . A composite score is also of interest, defined here as the sum of the two scale scores, $sc(Y) = sc(Y_1) + sc(Y_2)$. For this situation, the distribution of $sc(Y)$ is a function of the scale score distributions of $sc(Y_1)$ and $sc(Y_2)$ and also the joint distribution of $[sc(Y_1), sc(Y_2)]$. Specifically, the mean of $sc(Y)$ can be obtained as the sum of the means of both tests' scale scores, the variance of $sc(Y)$ is obtained as the sum of the variances of both tests' scale scores plus two times their covariance, and the skewness and higher moments reflect higher moments in $sc(Y_1)$ and $sc(Y_2)$ and in the joint distribution of $[sc(Y_1), sc(Y_2)]$. These relationships indicate that except for its mean, the distribution of the composite scale scores $sc(Y)$ is a function of the $sc(Y_1)$ and $sc(Y_2)$ scale scores that would be maintained through equatings of the Y_1 test forms to each other and the Y_2 test forms to each other, as well as the covariance and higher moments of the joint distributions of $[sc(Y_1), sc(Y_2)]$ that would not be maintained through the Y_1 equatings and the Y_2 equatings. Examples of composites produced from multiple tests

include those reported by ACT, SAT, and other programs. Kolen and Brennan (2014) summarized how a change in composite score standard deviation coincided with the change in intercorrelations of separate battery tests' scale scores and recommended specific checks for composite score comparability: "When composites are created for tests in a battery, it is important to check whether the composites are also comparable" (p. 425). Analogous results have been shown in linking situations involving latent composites (Dorans et al., 2014; Weeks, 2018). An implication is that composite score scales are best maintained directly rather than linked through indirect reliance on equated battery tests.

Predicting and Projecting

Predicting and projecting are the oldest and earliest examples of test linking (Holland & Dorans, 2006). Although these linkings are not symmetric like scale-aligning functions, various applications have been described in the psychometric literature. Predictions and projections have been recommended because of their increased prediction accuracy versus scale-aligning functions, such as when there are low correlations in the scale scores for which a concordance is desired (Dorans, 1999; 2004b), or in the IRT thetas of distinct measures intended to be linked (Holland & Hoskens, 2003; Schalet et al., 2021). Other uses are for establishing auxiliary scales, such as reports of normative growth using conditional norms (i.e., relative growth rather than the absolute growth intended to be shown in vertical scales). Conditional growth norms might be used to forecast how younger test takers obtaining specific scores on an earlier test are expected to perform on a later test (Betebenner, 2009; Castellano & Ho, 2013). For example, PSAT-to-SAT prediction tables are descriptions of expected growth at the student and school levels (Y. Kim et al., 2018a, 2018b). As described in Holland and Dorans's (2006) linking framework (also Linn, 1993; Mislevy, 1992), predictions and projections are the least restrictive type of linking in terms of requirements. They do, however, require a sample of test takers who took both tests. The usefulness and accuracy of prediction and projection tables are higher when the tables are applied to test takers who are most similar to the test takers used to produce the tables (Holland & Dorans, 2006). In other situations, the group dependences noted in prediction and projection tables can make them unstable, inaccurate, and "precarious" (Mislevy, 1992, p. 63), such as in applications to nonrepresentative groups or to groups over time (Thissen, 2007).

More Recent Linking Types

Since Holland and Dorans (2006) presented their test linking framework, different types of linking have been proposed and developed. Some prominent examples are summarized in this section and in Table 11.4. A focus of this discussion is the extent to which these more recently proposed linking types differ from those in the *Equating Requirements* section. Several of these proposals are based on desires to offer tests in more specific ways to test takers by offering tests in an increased number of administration modes (see the *Linking Tests Across Conditions of Measurement* section), supporting tests

with subjective scoring procedures (see the *Linking Subjectively Scored Tests* section), or developing the tests or the intended linking function tailored to specific test takers (see the *Local, Ability-Specific Test Linking*, and *Linking Adaptive Tests* sections). Additional proposals expand on methods for linking across nonequivalent groups (see the *Linking Tests Using Weak Anchors* section) and the *Linking State Tests to NAEP* section.

Linking Tests Across Conditions of Measurement

Testing programs might allow test takers to take their tests in different administration modes, such as translated into different languages, with modifications and accommodations for special populations, and in multiple delivery systems that include paper-pencil tests and computerized tests delivered online. When offering alternative editions of their tests that are administered under different conditions, the testing program can be faced with two choices. One choice is to ignore administration effects and treat the scores from a test given in different administrations in the same way. This choice could raise fairness concerns and result in an unknown degree of comparability in scores across measurement conditions (Pommerich, 2016). The other choice is to conduct linking studies to estimate and apply score adjustments that account for administration effects on scores (i.e., mode comparability studies). This choice may be possible for evaluating score effects from computerized versus paper testing modes, but is less feasible for tests administered with and without accommodations or in different languages where standard data collection designs may be less available (Thissen, 2016).

Mode comparability studies and linkings produced in these studies do not meet the equating requirement for equal conditions of measurement (see the *Equating Requirements* section) and are not expected to produce interchangeable scores. However, these studies might improve or at least inform the fairness of score reporting. Some reviews of paper-pencil versus computerized mode comparability studies suggest that scores from these modes are comparable more often than not (Pommerich, 2016), whereas other studies suggest that computerized testing can favor some subgroups over others (Kolen & Brennan, 2014). The mixed findings could be due to challenges in conducting mode comparability studies with traditional data collection designs (see the *Data Collection Designs* section) and the extent to which the administration modes differ.

Traditional data collection designs can be challenging to implement with mode comparability studies (Kolen, 2007). Mode comparability studies with randomly equivalent groups designs require that test takers be randomly assigned to an administration mode (paper-pencil, computerized, etc.), an assignment that can differ from the typical administrations of tests and modes by a testing program. A NEAT design implementation would require that test takers take an anchor test given in a single administration mode that may differ from the mode in which they take their actual test. Single group and counterbalanced designs can produce scores with order effects (Eignor, 2007). Traditional designs have been implemented in special studies for large-scale assessments like NAEP, the Programme for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS; Jia & Xi, 2021) and

have found that digitally administered items in mathematics and reading (NAEP) and math and science (TIMSS) were more difficult for fourth and eighth graders. Other studies have assessed mode effects in paper–pencil and online administrations of credentialing, licensure, and placement exams, using a variety of methods to account for test-taker nonequivalence and finding mixed results (Jones et al., 2022; S. Kim & Walker, 2021; Moses et al., 2021; Puhan & Kim, 2022).

A major question is how different the tests are when administered in each mode, such as in terms of their display of items and other test material, the test administration(s) (fixed, adaptive, etc.), and the extent of proctoring for the computerized administration (proctored at test centers vs. unproctored or online administrations). Discussions of these issues suggest that mode comparability linkings are most dissimilar with equatings that produce interchangeable scores across testing modes when the computerized test is adaptive and unlikely to exhibit equity with the paper–pencil scores (Eignor, 2007). In summarizing the transitions many admissions and placement programs made to online testing in efforts to continue administering tests in the COVID-19 pandemic, Camara (2020) noted that online administrations with especially large differences from paper–pencil tests risked the integrity, accuracy, reliability, validity, and fairness of the resulting scores and would “likely mean score trends and comparability cannot be maintained” (p. 13).

Altogether, the challenges with data collections and with potentially large differences in administration modes make mode comparability studies challenging in their implementation and their interpretation. Thissen (2016) provided a general discussion of testing and linking across different translations, grade levels, accommodations, and paper–pencil versus computerized modes, concluding that empirical investigations of fairness are needed for tests and linkings across conditions of measurement:

Continued vigilance is required. There is no way to guarantee that a use of a test is fair. All that can be done is to catalog carefully the ways in which comparisons could be unfair, and then check, with either statistics or reasoned judgment, whether unfairness exists. (Thissen, 2016, pp. 212–213)

Linking Subjectively Scored Tests

Tests that are composed of CR items present challenges for linking, such as tests that usually contain a small number of items and data collection designs that do not work the same way as for tests containing MC questions (Kolen & Brennan, 2014). The use of human raters to score tests containing a small number of CR questions can reduce reliability. Automated essay scoring approaches may avoid reliability reductions from human rater scoring, but they can also introduce their own challenges for score interpretability and fairness (see the *Ability Estimates From More Complex Models* section). The previously described issues about unreliable scores having less utility and producing linkings with less equity indicate that linkings for tests with CR items are not likely to produce interchangeable scores. Tests with small numbers of questions can also result

in inadequate coverage of the content domain, such that alternate forms may measure different constructs. Low reliability and inadequate content coverage can make equating and score interchangeability less likely for tests with CR questions.

When tests containing CR questions must be linked using a NEAT design, linking accuracy depends on a number of factors and procedures. A crucial issue with NEAT designs for CR tests is what to use for the anchor test. When the anchor contains CR questions, the scores from these anchors can reflect a particular type of bias due to leniency/stringency differences in human raters from the two administration groups (Tate, 1999, 2000; S. Kim et al., 2010a, 2010b). Trend scoring procedures are recommended for these biases, where the responses to a set of CR questions from a single group of test-takers are scored by raters in both administrations. A scale-aligning transformation for the sets of scores from these administrations is produced to account for the unintended differences in rater leniency/stringency across administrations and to produce scores that can be used more effectively as an anchor for linking the tests. The use of trend scored CR questions as an intact anchor, or as part of an anchor with MC question scores for mixed format tests, improves accuracy when linking mixed format tests with the NEAT design (S. Kim et al., 2010a, 2010b). Another choice available for mixed format tests is to use anchor tests composed only of MC items. This approach is most effective when the administration groups are similar to each other on the MC and CR sections, when the correlations of the MC and CR scores are high and similar across forms, and when the number of score points attributable to MC scores is large (Kolen & Lee, 2011, 2012, 2014, 2016, 2018).

Local, Ability-Specific Test Linking

If an equating satisfies equity, then it is a matter of indifference whether test takers at every given ability level take form X or Y (see the *Equating Requirements* section; Lord, 1980). Consider Lord's original definition of equity, defined over every ability level (θ value from an IRT model), and emphasize "each" ability level (van der Linden, 2011, p. 209; van der Linden, 2013, p. 262). A set of X -to- Y linking functions could be defined to satisfy this interpretation of equity at specific values of θ :

$$e_{Y\theta}(x) = G_{\theta}^{-1}[F_{\theta}(x)] \text{ such that } F_{\theta}[e_{Y\theta}(x)] = G_{\theta}[y] \text{ for each } \theta. \quad (36)$$

The goal of Equation 36 is to estimate each set of θ -conditional X and Y distributions (Lord & Wingersky, 1984) and use them to produce θ -conditional X -to- Y equipercentile functions. This proposal contrasts with the usual recommended practice to average the conditional distributions of X and of Y given θ (Equations 26 and 27) and conduct an equating for $e_{YT}(x)$ on the averaged distributions (Kolen & Brennan, 2014). If test takers' estimated θ s were available from an IRT model, a set of equating functions at each estimated θ value could be used to evaluate lack of equity in the equated scores, $e_{Y\theta}(x) - e_{YT}(x)$, rather than in distributions (Equation 18), analogous to evaluations of subpopulation invariance (Equation 20; Dorans & Holland, 2000).

One proposal is to use Equation 36 to produce and report θ -conditional scores to test takers, to satisfy equity and address equating bias attributable to θ -conditional measurement error (van der Linden, 2011, 2013). This proposal has been criticized for theoretical reasons and practical issues (Dorans, 2013; Holland, 2013). Some questions are discussed next to elaborate both perspectives, the proposal for θ -conditional procedures and counterarguments.

- Does the target population for equating, T , contain test takers at all θ values or is it θ -conditional? One perspective is that there are multiple θ -conditional target populations (van der Linden, 2011, p. 209). Another perspective defines T in terms of all test takers taking a test in a data collection design from Table 11.1 (i.e., test takers at all θ values). From the perspective that T contains test takers at all θ values, θ -conditional linking functions that differ from $e_{Y,T}(x)$ are an indication that the subpopulation invariance requirement is violated (Requirement e, Dorans, 2013). A θ -conditional interpretation would consider these differences to reflect different populations.
- Should administration conditions like scoring rules be applied in the same way to test takers at all θ s or should different scoring procedures be applied to test takers at different θ s? Unique scoring rules might be applied to test takers at specific θ values or consistently to test takers at all θ values. Note, however, that θ -conditional linking functions would imply that the equal measurement conditions requirement of equating is not satisfied (i.e., test takers with different θ values who take test X are scored differently from each other and likely from Y). Psychometricians have usually argued for the use of uniform scoring for all test takers taking a given test form (Livingston, 2004; Petersen, 2007).
- What are the implications of θ -conditional linking functions for score comparability? From one perspective, θ -conditional equipercentile functions correct for θ -conditional measurement errors (van der Linden, 2011, 2013). Another perspective is that θ -conditional equipercentile functions result in differential treatment of test takers (Dorans, 2013; Holland, 2013) and θ -conditional score interpretations that would be contradictory to interchangeability across different θ values (i.e., score comparisons across θ values are not supported).
- Is equating an adjustment for the estimated difficulties of X and Y in T or an adjustment for θ -conditional measurement error? One perspective is that the equating of observed scores involves adjusting for θ -conditional measurement errors (van der Linden, 2011, pp. 213, 223). The perspective of most other works cited in the *Equating* section is that equating involves adjusting test form scores for unintended differences in overall difficulty (see the *Equating Requirements section*), where the test forms involved are equally and highly reliable.

Rather than use Equation 36 to report θ -conditional linking functions to test takers, the recommendation in this chapter is to re-emphasize high reliability as a requirement for equating. This means that a single X -to- Y conversion applied to all test takers of X

would reflect less measurement error, would be a closer approximation of the equity requirement, and would have small differences from θ -conditional X -to- Y functions. The requirement for high and equal reliability in test forms being equated means Lord's (1980) theorem that only perfectly reliable true scores can be equated is more closely approximated. This perspective makes use of Equation 36 not as a score reporting strategy, but as one possible check for equating adequacy, namely, a check on the extent to which θ -conditional measurement error and (in)equity affect equating accuracy.

Linking Adaptive Tests

Adaptive testing is based on the premise that a test can be more precise when it is constructed so that its difficulty is matched ("tailored," Lord, 1980, p. 150) to test-taker ability. This matching of test-taker ability and test difficulty implies increased efficiency, in that a matched test can be shorter but measure at a precision that is similar to or higher than that of a longer test that may be too easy or too difficult for an unmatched test-taker group. Adaptive tests are usually IRT based in their scoring, linking, and scale scores, relying on a large pool of available test items with IRT statistics obtained from pretest and pre-equating data collections and using computer-based algorithms to administer the tests (van der Linden & Glas, 2010).

When the test is administered, a computer program implements sophisticated algorithms to obtain preliminary IRT-based ability estimates as a test taker takes test items. Then the computer program selects and administers subsequent test items with difficulties that match test takers' ability estimates. The computer might stop the test when the test taker's ability estimate reaches a predetermined level of precision. Additional procedures are needed to ensure that the item pool is adequately maintained, that items are not being overly selected or exposed so as to raise security concerns, that content specifications are met, and that items' IRT parameter estimates are not drifting during repeated use. Adaptive testing could be implemented at the item level (i.e., every subsequent item a test taker takes is based on his/her ability estimated from previously taken items). Adaptive testing could also be implemented at a small number of stages where test takers are routed to collections of items of differing difficulty (i.e., multistage tests, MSTs). Several examples of item-adaptive tests and MSTs are described by van der Linden and Glas (2010).

Computer-adaptive tests present several challenges for score linking. The adaptive tests are highly dependent on the item pools, such that systematic changes in the item pool through item reduction or the use of different exposure controls can result in test scores that are not interchangeable because of increased measurement error, violations of second-order equity, and changes in the distribution of reported scores (T. Wang & Kolen, 2001). Even when alternate item pools are built to be very similar, they might produce scores that are different enough to warrant additional linking beyond what is provided in the adaptive test (Segall, 1997). Adaptive tests rely on assumptions that scoring parameters are correct and item order and context effects are either minimal or

can be controlled. More thoughtfulness has been encouraged about these assumptions (Harris, 2023), and special research has been prompted to assess pretested items administered in varied positions (Davey & Lee, 2011). Importantly, the design of an adaptive MST will reflect one of several possible trade-offs of adaption and the extent to which content specifications are met:

More modules per stage may make a test more adaptable to a wider range of examinee proficiency levels, but then, more easy items and hard items are needed to build the MST modules. The items must also be selected to simultaneously meet all requisite content specifications for any test route taken by the examinee. This is often very challenging when the modules at a given stage must be matched on content and also span a fairly wide range of item difficulty. Similarly, fewer items per stage likewise encourages the use of more adaptation, but can result in routing decisions being made on smaller and smaller slices of the content domain. (Zenisky et al., 2010, p. 356)

Adaptive tests have some resemblances to vertical scales (see the *Vertical Scaling* section), in that they involve tests developed to differ in difficulty and possibly in content being administered to test takers of specific estimated ability. Both linking types can fail to meet the equating requirement for content and difficulty (Requirement a, see the *Equating Requirements* section) and both raise comparability challenges. Similar to how different vertical scaling results could be produced as students of different ages and/or grades take the tests being vertically scaled, different scores from computer-adaptive tests could be produced if test takers of higher and lower ability take one of several unique and adaptive tests differing in their difficulty and possibly content. These issues might be a matter of degree where, for example, simulations of MSTs that administer modules with relatively small differences in difficulty and content may not show large scoring errors in misrouted test takers (S. Kim & Moses, 2014). However, as the quote from Zenisky et al. (2010) suggests, it is possible that the adaption in adaptive testing could be more extreme, as is the case with item-level routing, such that tests targeted at specific ability levels differ from those that target other ability levels in their difficulty and content coverage. Greater degrees of adaption and larger differences in the adaptive tests administered to test takers can reduce the comparability of scores across test takers of different ability.

When unidimensional IRT models are employed, it is virtually impossible for test-taker scores to be successfully equated for adaptive “forms” because, by design, the adaptive routes differ in difficulty and indirectly modeled content, at least in part, as well as the test-taker groups that take the different routes. In effect, different ability groups take different test forms that are intentionally designed to be nonequivalent, which rules out equating as an attainable goal. (Similar statements apply to linking scores on paper–pencil and MST forms.) Just about all educational tests are multi-dimensional by design and by content specification. When a unidimensional IRT model is used and adaptations reflect ability, difficulty, and incompletely modeled

multidimensional content, statements about scores for test takers who take different test forms and routes being on the same scale are difficult to defend. Linkings of some kind may be possible, but the characteristics of any linked scores can be ascertained only through empirical studies that consider the accuracies of the IRT parameters and scores, the item pool, the adaptive settings, test forms for higher and lower ability test takers, etc. The strong claims supportable by an actual equating are not expected in such linkings.

Linking Tests Using Weak Anchors

Consider the situation where a linking is desired for tests X and Y , but the test data are not collected in one of the recommended data collection designs (Table 11.1). If these test data were obtained in a manner similar to a NEAT design in which X and Y were taken by nonequivalent groups, but with no anchor score data, what linking approaches might be available? One example of this situation occurs when the data from a paper-pencil test were intended to be collected in a randomly equivalent groups design, but test book spiraling procedures failed. Another situation occurs when the administration of a common anchor to both groups is not feasible because of security concerns. Finally, testing programs might intentionally introduce this type of design by offering their test in different administration modes and giving test takers the option to self-select into different test administration modes.

Two issues that determine the quality of linking are the similarity/dissimilarity of the administration groups and the representativeness of the available anchor(s), $A(s)$ for X and Y as indicated in the anchor, test correlations. An ideal situation occurs when the administration groups are samples from the same population. For this case, Holland and Dorans (2006) described the role of the anchor as one of reducing random variability and improving statistical precision, even when the anchor does not represent the tests:

When $P=Q$, the NEAT design is called an EG [randomly equivalent groups] design with anchor test. The two samples are drawn from a common population and the role of the anchor test changes. The anchor test becomes a covariate as in a randomized experimental design. It is used to gain precision in the estimation of the relevant parameters, rather than to adjust for group differences. For this special case, it is not necessary for A to measure the same construct that X and Y do, or even to be a test score. All that matters is for A to be correlated with both X and Y . When this is the case A is useful as a precision-increasing covariate. (p. 199)

For less than ideal situations where the populations of the administration groups cannot be considered equivalent, the anchor must reduce bias and variability (Holland & Dorans, 2006). The potential for greater linking error when administration groups are not equivalent means that it is especially important that the anchor be representative of, and correlated with, test scores. Although the use of inadequate anchors for a linking test administered to nonequivalent groups might improve linking accuracy, this use

may not sufficiently account for test-specific group differences or produce equatings that support interchangeable scores.

Several linking methods have been considered for situations where the administration groups differ and where multiple anchors and/or background variables might be available that could be incorporated into a test linking. Suggestions for using multiple scores were made by Angoff (1971) in a generalization of regression methods (Equation 17). Livingston et al. (1990) suggested propensity score matching as a way to incorporate multiple scores. Dorans and Wright (1993) showed that linking results obtained from matching administration groups based on a selection variable could be more accurate than methods based on matching groups with traditional anchors. Liou et al. (2001) described how demographic background variables might be used in test linking by making assumptions about the missing data and using corresponding imputation methods. Moses et al. (2010) showed how frequency estimation based on projected distributions (see the *Distinguishing Equating From Other Forms of Linking* section), categorized propensity scores, and missing data imputation could all be used to produce similar linking results when these methods were used with two anchor scores. Studies of categorized propensity scores based on background variables (Livingston, 2014; Wallin & Wiberg, 2019) and linear scale alignments with linear regressions of multiple background variables (Bränberg & Wiberg, 2011) showed that these approaches can increase precision and reduce variability in test linking. Statistical weighting procedures based on log-linear models have also been described for test linking based on background variables (Haberman, 2015) and on test takers' previous test scores (Y. Lee et al., 2019).

Most of these methods produce scale-aligning results that are similar to those obtained from projecting test score distributions for hypothetical administration groups defined by one or more anchors and/or background variables. It is possible that the linking results based on using an anchor and/or background variable have improved accuracy compared to an inappropriate randomly equivalent groups approach. However, to the extent that the anchors and/or background variables do not represent (and correlate highly with) the tests being linked, and also do not account for administration group differences, the linking results will likely have inadequacies, such as insufficiently controlled group differences that preclude interchangeable scores.

Linking State Tests to NAEP

The National Research Council published *Uncommon Measures* (Feuer et al., 1999) to address a debate in the late 1990s between those who favored voluntary national tests as a means of assessing the educational progress of students across the nation and those who believed that statistical linkages among existing tests could be used to achieve that purpose. The *Uncommon Measures* report examined the feasibility of linking the results of commercial and state tests, such as by linking these tests to each other and to the NAEP scales to support comparisons of students' achievement with national and international benchmarks and with students in other places. Feuer et al.

demonstrated that it was not possible to link state assessments for several reasons, including differences in the NAEP and state tests with respect to their content, format, margins of error, intended and actual uses of the tests, and the consequences attached to the test results.

Attempts to link state tests to the NAEP scale have continued since Feuer et al.'s (1999) report (Dorans, 2020), prompting discussions and questions about the interpretational value of the results. Thissen (2007, 2016) discussed an approach based on projection methods (Williams et al., 1998), and listed concerns about lack of population invariance, indications that linking results were not stable over time, study cost, and the unknown levels and differences in motivation for students taking a standardized test versus the NAEP test, which is more of a survey that lacks direct personal consequences for test takers. Other attempts involve linking state standards to NAEP scales using equipercentile methods (Braun & Qian, 2007; McLaughlin, 2000; McLaughlin & Bandeira de Mello, 2002). These attempts prompted concerns about whether the inferences and interpretations were too strong, about potential instabilities of the results over time (Koretz, 2007), and about comparisons that may not be defensible for the states and state tests that differ from NAEP with respect to content frameworks, implementation, and stakes (Ho & Haertel, 2007).

In another attempt to use the NAEP scale to connect disparate state assessment results, Reardon et al. (2021) reported linear scale-aligning linkings of school district means on state tests to the NAEP scale, based on statistically inferring the district means on the state tests from published passing rate distributions. Commentaries on Reardon et al.'s article by Bolt (2021), Davison (2021), Moses and Dorans (2021), and von Davier (2021) re-emphasized long-standing cautions about the large variation in blueprints used by different state tests and other differences with NAEP in terms of content, administration conditions, stakes, and test-taker motivation. Moses and Dorans (2021) provided empirical demonstrations that state-based linkings of district means from one test to another are not invariant across states, even when correlations in the district means of these scores are very high (.98 or higher). These discussions suggest that cross-state invariance evaluations are needed to support cross-state comparisons of districts.

DISCUSSION

The reporting scales for a large-scale testing program are the focus of five chapters of the previous editions of *Educational Measurement* (Angoff, 1971; Flanagan, 1951; Holland & Dorans, 2006; Kolen, 2006; Petersen et al., 1989). The current chapter updated these discussions. The *Scaling* section covered approaches to setting the scales for the test(s) of a large-scale testing program. The *Equating* section covered equating approaches for alternative versions of the same test. The *Linking* section discussed other linking approaches for relating the scores and scales of different tests. In these discussions, scaling, equating, and linking activities were described and related to a wider context

of testing activities, including test specifications, test form assembly, test scoring, test administration conditions (including data collection designs), and the intended purposes and uses of the test(s). When testing activities are consistent and tests are similar with respect to construction, administration, and purposes, the resulting scales are more likely to reflect their intended interpretations. When testing activities are inconsistent or are altered in tests that are scaled and linked to each other, the resulting scaling, linking, and equating procedures may not produce results that adequately support intended score interpretations.

Recent testing trends create and increase the challenges in establishing, maintaining, and relating the score scales of tests. Pursuits of alternative online administrations have increased since the COVID-19 pandemic, with mixed administration effects on reported scores (see the *Linking Tests Across Conditions of Measurement* section). There are also calls to revise testing in ways that address increasingly diverse testing populations, such as to develop and administer tests under less standardized conditions (Sireci, 2020), in ways that reflect test takers' sociocognitive backgrounds (Mislevy, 2018). In situations where testing programs compete for dwindling numbers of test takers, some programs have attempted to link their tests to other tests using substandard linking methodologies (Dorans & Moses, 2023). In other cases, unlinked scores are released with the presumption, but not necessarily the communication, that scores are linked based on nonempirical and untestable assumptions (Baldwin & Clauser, 2022; Dorans & Middleton, 2012). Assessment approaches can make different trade-offs of standardization, reliability, and comparability versus validity, group fidelity, and local uses (Brennan, 2006; Mislevy et al., 2025). Some of these trade-offs prompt recommendations to limit unwarranted score interpretations and broad comparisons of test takers that can be unfair (Dorans & Haberman, 2022; Moses, 2022, 2025).

As discussed in this chapter, large numbers of forms administered to smaller administration groups using weak data collection designs and little or no linking efforts can simultaneously inflate several types of error (random, systematic, violations of the equal construct requirement, etc.). They also raise concerns about the comparability of scores being released and used to make comparisons (Baldwin & Clauser, 2022; Moses, 2022). Experience indicates that these concerns have especially serious consequences for test score users when testing organizations are not forthcoming in communicating the procedures they use, the assumptions they make with their procedures, and the extent to which comparability in their scores is (un)supported. For these and other challenges, the recommended practices described in this chapter include checks to monitor and ensure that reported scales continue to reflect their intended interpretations. Additional recommendations are for checks on equating and linking results, especially those produced under limited and unique study conditions that may warrant interpretations that are restricted to the conditions of those linking studies rather than generalized to uses in broader testing populations (W. Lee & Brennan, 2021; Moses & Dorans, 2021; Moses, 2022).

The entity that produces reported scores has the ultimate responsibility for establishing and maintaining their reporting scales and for ensuring that the resulting scores and their interpretations are clearly communicated to test users (AERA et al., 2014). It is crucial that reported scores have clear, useful, and defensible interpretations and explanations based on sound scaling, equating, and (if necessary) linking, along with other supporting information. The ultimate goal is that reported scores, especially how they are produced and interpreted, should be integral and defensible evidence in support of validation arguments about test scores. It should be noted as well that this goal is unlikely to be attained without the active cooperation with and/or input from those who design, develop, administer, and market tests. Defensible scaling, equating, and linking is not simply the application of complex psychometrics to testing issues. All applied psychometrics involves assumptions, and the defensibility of results rests heavily on the credibility of these assumptions in the specific testing context under consideration, with appropriate attention given to quantifying and communicating likely error in reported scores:

Whether scores are equated, linked in some weaker sense, or rescaled, however, the overarching consideration in my opinion is that users be given appropriate guidance about score interpretation and use. Part of that guidance ought to be explicit indications of the amount of error in scores and in the likely uses made of scores, as well as admonitions about likely misinterpretations of scores. (Brennan, 2007, p. 175)

ACKNOWLEDGMENTS

My chapter was greatly improved by reviews from my reviewer-collaborators, Robert Brennan, Neil Dorans, and Michael Kolen, an early review from Derek Briggs, helpful discussions with Jonathan Weeks, and suggestions from editors Linda Cook and Mary Pitoniak. I am very thankful for these reviews. I am also grateful for the collaborations with all of my coauthors cited throughout this chapter.

REFERENCES

- ACT. (2001). *EXPLORE technical manual*.
- Allen, N. L., Carson, J. E., & Zelenak, C. A. (1999) *The NAEP 1996 technical report*. National Center for Education Statistics.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Andrich, D. (1988). *Rasch models for measurement*. Sage Publications.
- Angoff, W. H. (1962). *The equating of nonparallel tests* (ETS Research Memorandum No. RM-62-02). ETS.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.

- Angoff, W. H., & Cowell, W. R. (1986). An examination of the assumption that the equating of parallel forms is population-independent. *Journal of Educational Measurement*, 23, 327–345.
- Baldwin, P., & Clauser, B. (2022). Historical perspectives on score comparability issues raised by innovations in testing. *Journal of Educational Measurement*, 59, 140–160.
- Becker, K. A. (2003). *History of the Stanford–Binet intelligence scales: Content and psychometrics* (Stanford–Binet intelligence scales, 5th Ed., Assessment Service Bulletin No. 1). Riverside.
- Betebenner, D. W. (2009). *Growth, standards and accountability*. Colorado Department of Education.
- Binet, A., & Simon, T. (1916). *The development of intelligence in children (the Binet–Simon Scale)*. Williams & Wilkins Company.
- Bolt, D. (2021). Commentary on Reardon, Kalorgrides, and Ho’s “Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale.” *Journal of Educational and Behavioral Statistics*, 46(2), 168–172.
- Bränberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement*, 48, 419–440.
- Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). Academic Press.
- Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199–216). Springer Science + Business Media.
- Brennan, R. L. (Ed.). (1989). *Methodology used in scaling the ACT assessment and P-ACT+*. CASMA.
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
- Brennan, R. L. (2004). *Manual for LEGS Version 2.0*. (3). CASMA.
- Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), pp. 1–16). Praeger.
- Brennan, R. L. (2007). Tests in transition: Discussion and synthesis. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 161–175). Springer-Verlag.
- Brennan, R. L. (2010). *First-order and second-order equity in equating* (CASMA Research Report No. 30). CASMA.
- Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes* (CASMA Monograph No. 1). Center for Advanced Studies in Measurement and Assessment, The University of Iowa.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50, 204–226.
- Briggs, D. C. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge.

- Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, 38, 551–576.
- Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice*, 28, 3–14.
- Broussard, M. (2020, September 8). When algorithms give real students imaginary grades. *New York Times*. <https://www.nytimes.com/2020/09/08/opinion/international-baccalaureate-algorithm-grades.html>
- Camara, W. (2020). Never let a crisis go to waste: Large-scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practice*, 39, 10–18.
- Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. Longmans, Green.
- Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Council of Chief State School Officers.
- Chen, H. H., Livingston, S. A., & Holland, P. W. (2011). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 185–200). Springer.
- College Board. (n.d.). *About AP scores*. Retrieved January 1, 2025 from <https://apstudents.collegeboard.org/about-ap-scores>
- College Board/ACT. (2018). *Guide to the 2018 ACT/SAT concordance*. <https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf>
- Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Prentice Hall.
- Cronbach, L. J. (1949). *Essentials of psychological testing*. Harper & Brothers.
- CTB/McGraw-Hill. (2001). *TerraNova technical report*.
- Cui, Z., & Kolen, M. J. (2009). Evaluation of two new smoothing methods in equating: The cubic b-spline presmoothing method and the direct presmoothing method. *Journal of Educational Measurement*, 46, 135–158.
- Davey, T., & Lee, Y.-H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised General Test* (ETS Research Report No. RR-11-26). ETS.
- Davison, M. (2021). Commentary on “Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale.” *Journal of Educational and Behavioral Statistics*, 46(2), 172–186.
- Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (College Board Research Report No. 99-1). College Board.
- Dorans, N. J. (2002). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement*, 39, 59–84.
- Dorans, N. J. (2004a). Assessing the population sensitivity of equating functions. [Special issue]. *Journal of Educational Measurement*, 41(1).
- Dorans, N. J. (2004b). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28, 227–246.

- Dorans, N. J. (2011). Holland's advice for the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). Springer.
- Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, 31(4), 20–37.
- Dorans, N. J. (2013). On attempting to do what Lord said was impossible: Commentary on van der Linden's "Some conceptual issues in observed-score equating." *Journal of Educational Measurement*, 50, 304–314.
- Dorans, N. J. (2018). Scores, scales, and score linking. In P. Irwing, T. Booth, & D. J. Hughs (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 573–605). Wiley Blackwell.
- Dorans, N. J. (2020). Uncommon measures revisited (ETS Research Report No. RR-20-04). ETS.
- Dorans, N. J. (2022, April 21–26). *Revisiting lingering linking issues: Now; Never-ending?* (Invited 2021 Robert L. Linn Distinguished Address). American Educational Research Association Annual Meeting, San Diego, CA, United States.
- Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. In I. M. Lawrence, N. J. Dorans, M. D. Feigenbaum, N. J. Feryok, A. P. Schmitt, & N. K. Wright (Eds.), *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (ETS Research Memorandum No. RM-94-10, pp. 91–122). ETS.
- Dorans, N. J., & Haberman, S. J. (2022). Recent challenges to maintaining score comparability: A commentary. *Journal of Educational Measurement*, 59, 251–264.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281–306.
- Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, 3(3), 245–254.
- Dorans, N. J., Liang, L., & Puhan, G. (2010). *Aligning scales of certification tests* (ETS Research Report No. 10-07). ETS.
- Dorans, N. J., Lin, P., Wang, W., & Yao, L. (2014). *The invariance of latent and observed linking functions in the presence of multiple latent test-taker dimensions* (ETS Research Report No. RR-14-41). ETS. <http://dx.doi.org/10.1002/ets2.12041>
- Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, 73, 24–34.
- Dorans, N. J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, 49, 1–18.
- Dorans, N. J., & Moses, T. (2023) Score equating: An aspirational form of score linking. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (vol. 14, 4th ed., pp. 236–248). Elsevier.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: Toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 21–42). Springer.

- Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. Springer-Verlag.
- Dorans, N. J., & Puhon, G. (2017). Contributions to score linking theory and practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment* (pp. 79–132). ETS.
- Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 179–198). Springer-Verlag.
- Dorans, N. J., & Wright, N. K. (1993). *Using the selection variable for matching or equating* (Research Report No. RR-93-04). ETS.
- Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). Praeger.
- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to support army selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.
- ETS. (2005). *The Comprehensive English Language Learning Assessment: Technical report*.
- Eignor, D. R. (2007). Linking scores derived under different modes of test administration. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 135–159). Springer-Verlag.
- Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist*, 61, 50–55.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Engelhard, G. (1984). Thorndike, Thurstone, and Rasch: A comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, 8, 21–38.
- Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments* (pp. 52–59). Routledge.
- Ferguson, A., Myers, C., Bartlett, R., Banister, H., Bartlett, F. C., Brown, W., Campbell, N., Craik, K., Drever, J., Guild, J., Houstoun, R., Irwin, J., Kaye, G., Philpott, S., Richardson, L., Shaxby, J., Smith, T., Thouless, R., & Tucker, W. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, 1, 331–349.
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). National Academies Press.
- Flanagan, J. C. (1939). *The cooperative achievement tests. A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores*. American Council on Education Cooperative Test Service.
- Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). American Council on Education.

- Forsyth, R. A. (1991). Do NAEP scales yield criterion-referenced interpretations? *Educational Measurement: Issues & Practice*, 10(3), 3–9, 16.
- Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and square root. *Annals of Mathematical Statistics*, 21, 607–611.
- Gao, R., He, W., & Ruan, C. (2012). *Does preequating work? An investigation into a preequated testlet-based college placement exam using postadministration data* (ETS Research Report No. RR-12-12). ETS. <https://doi.org/10.1002/j.2333-8504.2012.tb02294.x>
- Golub-Smith, M. L., & Moses, T. P. (2014). How the scales for the GRE revised test were defined. In C. Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE revised General Test: A compendium of studies* (pp. 2.2.1–2.2.9). ETS.
- Gonzalez, J., & Wiberg, M. (2017). *Applying test equating methods: Using R (Methodology of Educational Measurement and Assessment)*. Springer.
- Gulliksen, H. (1950). *Theory of mental tests*. Wiley.
- Guo, H., Liu, J., Curley, E., Dorans, N., & Feigenbaum, M. (2012). *The stability of the score scale for the SAT Reasoning Test from 2005–2012* (ETS Research Report No. RR-12-15). ETS.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139–150.
- Haberman, S. (2011). Using exponential families for equating. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 125–140). Springer.
- Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, 40, 254–273.
- Haberman, S., & Dorans, N. J. (2011). *Sources of scale score inconsistency* (ETS Research Report No. RR-11-10). ETS.
- Haberman, S. J., Sinharay, S., & Lee, Y. (2011). Statistical procedures to evaluate quality of scale anchoring (ETS Research Report No. RR-11-04). ETS.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 65–110). Praeger.
- Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Stanford University, Applied Mathematics and Statistics Laboratory.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). American Council on Education/Praeger.
- Hanson, B. A. (1991). A note on Levine's formula for equating unequally reliable tests using data from the common item nonequivalent groups design. *Journal of Educational Statistics*, 16, 93–100.
- Harris, D. J. (2023). NCME presidential address: Some musings on comparable scores, *Educational Measurement: Issues and Practice*, 43, 6–15.
- Harris, D. J., & Kolen, M. J. (1986). Effect of examinee group on equating relationships. *Applied Psychological Measurement*, 10, 35–43.

- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40, 685–711.
- Ho, A. D., & Haertel, E. H. (2007). *(Over)-interpreting mappings of state performance standards onto the NAEP scale* [Commissioned paper]. Council of Chief State School Officers.
- Hölder, O. (1901). Die Axiome der Quantität und die Lehre vom Mass [The axioms of quantity and the theory of measurement]. *Berichte über die Verhandlungen der Königlich Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, 53, 1–46.
- Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. *Proceedings of the Social Statistics Section of the American Statistical Association*, pp. 27–29.
- Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 7–30). Springer-Verlag.
- Holland, P. W. (2008, March 3–5). *The first four generations of test theory* [Paper presentation]. Association of Test Publishers Conference on Innovations in Testing, Dallas, TX, United States.
- Holland, P. W. (2013). Comments on van der Linden's critique and proposal for equating. *Journal of Educational Measurement*, 50, 286–294.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger.
- Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, 68, 123–149.
- Holland, P. W., & Rubin, D. B. (Ed.). (1982). *Test equating*. Academic.
- Holland, P. W., & Thayer, D. T. (1989). *The kernel method of equating score distributions* (ETS Technical Report No. RR-89–84). ETS.
- Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, 25, 133–183.
- Holland, P. W., & Wightman, L. E. (1982). Section pre-equating: A preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. (pp. 271–297). Academic Press.
- Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practice*, 3, 8–14.
- Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa tests: Guide to research and development*. Riverside Publishing.
- Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5, e208. <https://doi.org/10.7717/peerj-cs.208>

- Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion referenced interpretation. *Journal of Educational & Behavioral Statistics*, 23, 35–56.
- Iowa Tests of Educational Development. (1958). *Manual for school administrators*. 1958 revision. State University of Iowa.
- Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. *Journal of Educational Statistics*, 10, 143–160.
- Jia, Y., & Xi, N. (2021, June 9–11). *Design considerations for linking large scale survey assessments across modes*. National Council of Measurement in Education Annual Meeting.
- Jones, P., Tong, Y., Liu, J., Borglum, J., & Primoli, V. (2022). Score comparability between online proctored and in-person credentialing exams. *Journal of Educational Measurement*, 59, 180–207.
- Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge.
- Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2, 389–423.
- Kelley, T. L. (1914). Comparable measures. *Journal of Educational Psychology*, 5, 589–595.
- Kelley, T. L. (1923). *Statistical methods*. Macmillan.
- Kim, H., Brennan, R. L., & Lee, W. C. (2017). Structural zeros and their implications with log-linear bivariate presmoothing under the internal anchor design. *Journal of Educational Measurement*, 54(2), 145–164.
- Kim, S. Y., Kim, Y., & Moses, T. (2020). *Impact of degrees of postsmoothing on long-term equated scale score accuracy* (CASMA Research Report No. 54). CASMA. <http://www.education.uiowa.edu/casma>
- Kim, S., & Moses, T. (2014). *An investigation of the impact of misrouting under two-stage multistage testing: A simulation study* (ETS Research Report No. RR-14-01). ETS. <http://dx.doi.org/10.1002/ets2.12000>.
- Kim, S., & Moses, T. (2025). Item response theory (IRT) proficiency estimation methods under multistage testing. In D. Yan (Ed.), *Research for practical issues and solutions in computerized multistage testing* (Vol. 1, pp. 185–199). ETS.
- Kim, S., & Walker, M. (2021). *Assessing mode effects of at home testing without a randomized trial* (ETS Research Report No. RR-21-10). ETS. <https://doi.org/10.1002/ets2.12323>
- Kim, S., Walker, M. E., & McHale, F. (2010a). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, 46(1), 36–53.

- Kim, S., Walker, M. E., & McHale, F. (2010b). Investigating the effectiveness of equating designs for constructed-response tests in large-scale assessments. *Journal of Educational Measurement*, 47, 186–201.
- Kim, Y. K., Moses, T., Kolen, M. J., & Hendrickson, A. (2016). *Scaling for the SAT Suite of Assessments*. College Board.
- Kim, Y. K., Moses, T., & Zhang, X. (2018a). *Student-level growth estimates for the SAT® suite of assessments*. College Board. <https://collegereadiness.collegeboard.org/pdf/student-level-sat-suite-growth-estimates.pdf>
- Kim, Y. K., Moses, T., & Zhang, X. (2018b). *School-level growth estimates for the SAT® suite of assessments*. College Board. <https://collegereadiness.collegeboard.org/pdf/sat-suite-growth-estimates.pdf>
- Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercetile equating. *Journal of Educational Statistics*, 9, 25–44.
- Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement*, 25, 97–110.
- Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement*, 41, 3–14.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Praeger.
- Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). Springer-Verlag.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer Publishing.
- Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 29(4), 285–307.
- Kolen, M. J., & Lee, W.-C. (2011). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 1). CASMA. <https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.1.pdf>
- Kolen, M. J., & Lee, W.-C. (2012). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 2). CASMA. <https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.2.pdf>
- Kolen, M. J., & Lee, W.-C. (2014). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 3). CASMA. <https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.3.pdf>
- Kolen, M. J., & Lee, W.-C. (2016). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 4). CASMA. <https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.4.pdf>
- Kolen, M. J., & Lee, W.-C. (2018). *Mixed-format tests: Psychometric properties with a primary focus on equating* (Vol. 5). CASMA. <https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-2.5.pdf>

- Kolen, M. J., Tong, Y., & Brennan, R. L. (2011). Scoring and scaling educational tests. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 43–58). Springer.
- Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339–353). Springer-Verlag.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement, Vol. 1: Additive and Polynomial Representations*. Academic Press.
- Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, 18, 89–109.
- Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, 64(3), 478–497.
- Lee, W., & Brennan, R. L. (2021). *An independent review of Educational Testing Service's study on predictive validity of the Graduate Record Exams for making law school admissions decisions*. https://www.americanbar.org/content/dam/aba/administrative/legal_education_and_admissions_to_the_bar/2021/20210907-casma-report-final.pdf
- Lee, Y., & Haberman, S. H. (2013). Harmonic regression and scale stability. *Psychometrika*, 78, 815–829.
- Lee, Y., Haberman, S., & Dorans, N. J. (2019). Use of adjustment by minimum discriminant information in linking constructed-response test scores in the absence of common items. *Journal of Educational Measurement*, 56, 452–472.
- Lee, Y., & von Davier, A. A. (2011). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 159–173). Springer.
- Levine, R. S. (1955). *Equating the score scales of alternate forms administered to samples of different ability* (ETS Research Bulletin No. RB-55-23). ETS.
- Li, D. (2023). *Investigating the impact of equating on score variability using generalizability theory*. IMPS Presentation.
- Lindquist, E. F. (1953). Selecting appropriate score scales for tests. Discussion. In *Proceedings of the 1952 Invitational Conference on Testing Problems* (pp. 34–40). ETS.
- Lindquist, E. F. (1964, February 12–15). *Equating scores on non-parallel tests* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83–102.
- Liou, M., & Cheng, P. E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics*, 20, 259–286.
- Liou, M., Cheng, P. E., & Li, M. Y. (2001). Estimating comparable scores using surrogate variables. *Applied Psychological Measurement*, 25, 197–207.

- Liu, C., & Kolen, M. J. (2018). A comparison of strategies for smoothing parameter selection for mixed-format tests under the random groups design. *Journal of Educational Measurement*, 55, 564–581.
- Liu, J., & Dorans, N. J. (2012). Assessing the practical equivalence of conversions when measurement conditions change. *Journal of Educational Measurement*, 49, 101–115.
- Liu, J., & Dorans, N. J. (2013). Assessing a critical aspect of construct continuity when test specifications change or test forms deviate from specifications. *Educational Measurement: Issues and Practice*, 32(1), 15–22.
- Liu, J., Dorans, N. J., & Moses, T. (2010, May 1–3). *Evaluating the subpopulation sensitivity of the ACT–SAT concordances* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Denver, CO, United States.
- Liu, J., Guo, H., & Dorans, N. J. (2014). *A comparison of raw-to-scale conversion consistency between single- and multiple-linking using a nonequivalent groups anchor test design* (Research Report No. RR-14-13). ETS.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. ETS.
- Livingston, S. A. (2014). *Demographically adjusted groups for equating test scores* (ETS Research Report No. RR-14-30). ETS. <https://doi.org/10.1002/ets2.12030>
- Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, 3, 73–95.
- Livingston, S. A., & Kim, S. (2011). New approaches to equating with small samples. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 109–122). Springer.
- Lord, F. M. (1950). *Notes on comparable scales for test scores* (ETS Research Bulletin No. RB-50-48). ETS.
- Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8, 750–751.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.
- Lord, F. M. (1982). The standard error of equipercentile equating. *Journal of Educational Statistics*, 7, 165–174.
- Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233–245.
- Lord, F. M. (1985). Estimating the imputed social cost of errors of measurement. *Psychometrika*, 50, 57–68.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23(2), 157–162.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison–Wesley.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score “equatings.” *Applied Psychological Measurement*, 8, 452–461.
- Lorge, I. (1951). The fundamental nature of measurement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 533–559). American Council on Education.

- Lottridge, S., Woolf, S., Young, M., Jafari, A., & Ormerod, C. (2023). The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning*, 39, 787–803.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Luecht, R., & Burke, M. (2020). Reconceptualizing items: From clones and automatic item generation to task model families. In H. Jiao & R. W. Lissitz (Eds.), *Application of artificial intelligence to assessment* (pp. 25–49). Information Age Publishing.
- Marco, G. L., & Abdel-Fattah, A. A. (1991). Developing concordance tables for scores on the Enhanced ACT Assessment and the SAT. *College & University*, 66, 187–194.
- McLaughlin, D. H. (2000). *Protecting state NAEP trends from changes in SD/LEP inclusion rates* (Report to the National Institute of Statistical Sciences). American Institutes for Research.
- McLaughlin, D. H., & Bandeira de Mello, V. (2002, April 1–5). *Comparison of state elementary school mathematics achievement standards using NAEP 2000* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States.
- Michell, J. (2008). Is psychometrics pathological science? *Measurement*, 6, 7–24.
- Mislevy, R. J. (1992, December). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Center Report). ETS.
- Mislevy, R. J. (2012). Comments on Neil Dorans's NCME Career Award Address: The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement: Issues and Practice*, 31(4), 38–39.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J., Olivari, M. E., Slomp, D., Wolf, A. C. E., & Elliot, N. (2025). An evidentiary-reasoning lens for socioculturally responsive assessment. In R. Bennett, L. Darling-Hammond, & A. Badrinarayan, (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. (pp. 199–241). Routledge.
- Modu, C. C., & Stern, J. (1975). *The stability of the SAT score scale* (ETS Research Bulletin No. RB-75-9). ETS.
- Morris, C. N. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). Academic Press.
- Moses, T. (2008). Using the kernel method of test equating for estimating the standard errors of population invariance measures. *Journal of Educational and Behavioral Statistics*, 33, 137–157.
- Moses, T. P. (2011). Log-linear models as smooth operators: Holland's statistical applications and their practical uses. In N. J. Dorans & S. Sinharay (Eds.), *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 185–202). Springer.
- Moses, T. (2014a). Quantifying error and uncertainty reductions in scaling functions: An ITEMS module. *Educational Measurement: Issues and Practice*, 33, 29–40.
- Moses, T. (2014b). Alternative smoothing and scaling strategies for weighted composite scores. *Educational and Psychological Measurement*, 74(3), 516–536.

- Moses, T. (2022). Linking and comparability across conditions of measurement. Established frameworks and proposed updates. *Journal of Educational Measurement*, 59, 231–250.
- Moses, T. (2025). Linking responsive assessments: Challenges and possibilities. In R. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy*. (pp. 180–198). Routledge.
- Moses, T., Deng, W., & Zhang, Y. L. (2010). *The use of two anchors in the nonequivalent groups with anchor test (NEAT) equating* (ETS Research Report No. RR-10-23). ETS.
- Moses, T., & Dorans, N. J. (2021). Aggregate-level test-scale linking: A new solution for an old problem? *Journal of Educational and Behavioral Statistics*, 46(2), 187–202.
- Moses, T. P., Dorans, N. J., Miao, J., & Yon, H. (2006). *Weighting strategies for Advanced Placement exams*. ETS.
- Moses, T., & Golub-Smith, M. (2011). *A scaling method that produces scale score distributions with specific skewness and kurtosis* (ETS Research Memorandum No. RM-11-04). ETS.
- Moses, T. P., & Holland, P. W. (2008). *Notes on the general framework for observed score equating* (ETS Research Report No. RR-08-59). ETS.
- Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate loglinear models. *British Journal of Mathematical and Statistical Psychology*, 63, 557–574.
- Moses, T., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (ETS Research Report No. RR-07-16). ETS.
- Moses, T., & Kim, S. (2015). Methods for evaluating composite reliability, classification consistency, and classification accuracy for mixed-format licensure tests. *Applied Psychological Measurement*, 39, 314–329.
- Moses, T., & Kim, Y. K. (2017). Stabilizing conditional standard errors of measurement in scale score transformations. *Journal of Educational Measurement*, 54, 184–199.
- Moses, T., Yang, W., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement*, 44, 157–178.
- Moses, T. P., & Zhang, W. (2011). Standard errors of equating differences: Prior developments, extensions and simulations. *Journal of Educational and Behavioral Statistics*, 36, 779–803.
- Moses, T., Kim, Y. K., Duffy, L., Lee, D., Patel, P., & Kulscar, B. (2021). *Linking analyses and results for digital tests in the AP 2021 administration*. [Unpublished documentation]. College Board.
- Mroch, A. A., Suh, Y., Kane, M. T., & Ripkey, D. R. (2009). An evaluation of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research and Perspectives*, 7(3–4), 174–193.
- Ogasawara, H. (2001). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics*, 26, 31–50.

- Ogasawara, H. (2003). Asymptotic standard errors of IRT observed-score equating methods. *Psychometrika*, 68, 193–211.
- Otis, A. S. (1922). The method for finding the correspondence between scores in two tests. *Journal of Educational Psychology*, 13, 529–545.
- Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. National Academies Press.
- Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237–255.
- Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). Springer-Verlag.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Macmillan.
- Phillips, S. E. (2016). Legal aspects of test fairness. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement: Vol 3. NCME application of educational assessment and measurement* (pp. 239–266). Routledge.
- Pommerich, M. (2007). Concordance: The good, the bad, and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 199–216). Springer-Verlag.
- Pommerich, M. (2016). The fairness of comparing test scores across different tests or modes of administration. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement: Vol 3. NCME applications of educational assessment and measurement* (pp. 111–134). Routledge.
- Puhan, G., & Kim, S. (2022). Score comparability issues with at-home testing and how to address them. *Journal of Educational Measurement*, 59, 161–179.
- Puhan, G., Moses, T., Grant, M., & McHale, F. (2009). Small-sample equating using a single-group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement*, 46, 344–362.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute Educational Research.
- Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district distributions to a common scale. *Journal of Educational and Behavioral Statistics*, 46, 138–167.
- Rijmen, F., Manalo, J., & von Davier, A. A. (2009). Asymptotic and sampling-based standard errors for two population invariance measures in the linear equating case. *Applied Psychological Measurement*, 33, 222–237.
- Schalet, B. D., Lim, S., Cella, D., & Choi, S. W. (2021). Linking scores with patient-reported health outcome instruments: A validation study and comparison of three linking methods. *Psychometrika*, 86(3), 717–746.
- Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). American Psychological Association.

- Sireci, S. G. (2020). Standardization and UNDERSTANDARDization in educational assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105.
- Sinharay, S., & Holland, P. W. (2006a). *Choice of anchor test in equating* (ETS Research Report No. RR-06-35). ETS.
- Sinharay, S., & Holland, P. W. (2006b). *The correlation between the scores of a test and an anchor test* (ETS Research Report No. RR-06-04). ETS.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29(3), 184–203.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 1–76). John Wiley.
- Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, 36(4), 336–346.
- Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, 37, 329–346.
- Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–310). Springer-Verlag.
- Thissen, D. (2016). Chapter 11: Commentary on the assessment of the fairness of comparisons under divergent measurement conditions. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 203–214). Routledge.
- Thissen, D., Liu, Y., Magnus, B., & Quinn, H. (2015). Extending the use of multidimensional IRT calibration as projection: Many-to-one linking and linear computation of projected scores. In L. van der Ark, D. Bolt, W. C. Wang, J. Douglas, & S. M. Chow (Eds.), *Quantitative psychology research: Springer Proceedings in Mathematics & Statistics* (Vol. 140, pp. 1–16). Springer.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum Associates.
- Thorndike, E. L. (1910). The measurement of the quality of handwriting. *Teachers College Record*, 11, 86–151.
- Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements*. Columbia University, Teachers College.
- Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology*, 6, 29–33.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16(7), 433–451.

- Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, 17, 446–457.
- Thurstone, L. L. (1927). The law of comparative judgement. *Psychological Review*, 34, 278–286.
- Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*. No. 1. University of Chicago Press.
- van der Linden, W. J. (2011). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 201–223). Springer.
- van der Linden, W. J. (2013). Some conceptual issues in observed-score equating. *Journal of Educational Measurement*, 50, 249–285.
- van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer.
- von Davier, A. A. (Ed.). (2011). *Statistical models for test equating, scaling, and linking*. Springer.
- von Davier, A. A. (2021). Commentary on “Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale.” *Journal of Educational and Behavioral Statistics*, 46(2), 203–208.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004a). *The kernel method of test equating*. Springer Publishing.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004b). The chain and post-stratification methods of observed-score equating: Their relationship to population invariance. *Journal of Educational Measurement*, 41, 15–32.
- von Davier, A. A., & Liu, M. (2007). Population invariance. [Special issue]. *Applied Psychological Measurement*, 32(1).
- Wallin, G., & Wiberg, M. (2019). Kernel equating using propensity scores for nonequivalent groups. *Journal of Educational and Behavioral Statistics*, 44, 390–414.
- Wang, J., & Smith, R. (2003). *NYSESLAT calibration and linking summary*. ETS.
- Wang, T. (2011). An alternative continuization method: The continuized log-linear method. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 141–157). Springer.
- Wang, T., & Brennan, R. L. (2008). A modified frequency estimation equating method for the common-item nonequivalent groups design. *Applied Psychological Measurement*, 33, 118–132.
- Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement*, 38, 19–49.
- Weeks, J. (2018). An application of multidimensional vertical scaling. *Measurement: Interdisciplinary Research and Perspectives*, 16, 139–154.
- Wendler, C., & Bridgeman, B. (2014). *The Research Foundation for the GRE (R) revised General Test: A compendium of studies*. ETS.
- Williams, V. S. L., Rosa, K. R., McLeod, L. D., Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale: Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement*, 35, 277–296.

- Wise, L. L. (2017). Commentary I: Validation of score meaning in the next generation of assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 52–59). Routledge.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116.
- Wright, B. D. (1994). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions*, 6, 196–200.
- Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 273–283). Springer-Verlag.
- Yen, W. M., & Fitzpatrick, A. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Praeger.
- Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement*, 28, 274–289.
- Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. Holt.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer.
- Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice*, 42, 44–58.

NOTE

1. The variance of mean-score differences plus the relative error variance in classical test theory is the absolute error variance in generalizability theory (Brennan, 2001).

Standard Setting

A Cognitive and Social Model

Steve Ferrara
HumRRO

Susan Davis-Becker
ACS Ventures

Priya Kannan
WestEd

Katherine Reynolds
Boston College

Standard setting is policymaking. Educational performance standards represent the aspirations that a governing entity holds for the learning and educational achievement of its school-age children. Credentialing performance standards represent the minimal level of competence necessary to ensure that professionals can provide safe and effective practice. In both situations, Cizek and Bunch (2007) noted that standard setting is “a procedure that enables participants . . . to bring to bear their judgments . . . to translate the policy positions of authorizing entities into locations on a score scale” (p. 19).

Standard setting is also a technology. It is a set of concepts, tools, methods, and procedures that we implement to create reasonable and defensible performance standards. As Reckase (2006) described it, standard-setting procedures are “mechanism[s] for helping panelists translate their intentions to the score scale” (p. 5, see also Reckase, 2023). In educational testing, performance standards are composed of a test (or test design), performance levels and descriptors, and corresponding cut scores. In credentialing testing, performance standards are composed of a test (or test design), defined expectations for minimum competency in a professional role (i.e., knowledge, skills, abilities), and a cut score. Most research in standard setting has focused on the components of this technology, including comparing methods, investigating components of individual methods, and exploring participant experiences.

Standard setting has been defined as the “level of performance on a test [that] is required for a test taker to be classified into a given performance category . . . [and] the process of deriving those levels” (Cizek, 2012b, p. 4) and, more specifically, “the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance” (Cizek, 1993, p. 100). Performance standards that are created from a standard-setting process define intended, appropriate “interpretations of test scores [and other information] for the intended uses” (American Educational Research Association [AERA] et al., 2014, p. 1). For example, performance-level labels like *Proficient* and *Advanced* and their associated descriptors guide interpretation of individual and group performance on state educational summative assessments. Candidates who achieve a “pass” on a credentialing exam have satisfied the knowledge and skill requirements for the license, certification, or certificate. These interpretations of test performance, based on performance standards, are claims that require validation.

“Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other” (Cizek, 2001, p. 5). The key concepts in these views on standard setting—policymaking and translating policy intentions to a test score scale, levels of performance, and a technology with prescribed concepts, tools, and procedures—highlight elements of standard setting that we address, to varying degrees, in this chapter. We focus especially in this chapter on the role and influences of human judgment and decision-making in standard setting.

Most standard settings include a workshop where experts (i.e., panelists), working independently and collectively to achieve a shared goal (i.e., usually convergence,

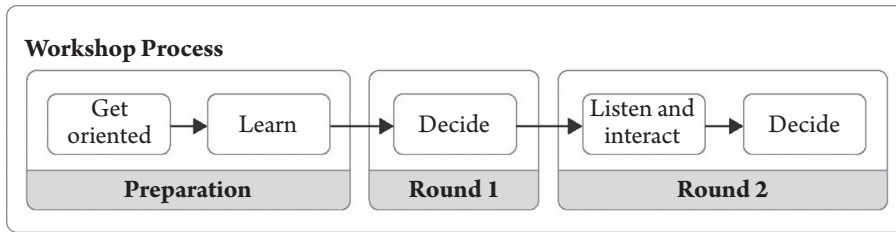


FIGURE 12.1
Typical Standard Setting Workshop Process

not consensus), are guided by a workshop facilitator through a method and set of procedures. Numerous methods exist for guiding a standard setting, and a typical process is shown in Figure 12.1. This begins by preparing the panelists to participate in the standard setting through orientation and training led by an expert facilitator. The process then moves into multiple rounds of standard-setting judgments where panelists can apply the standard-setting method (Round 1), review feedback as to how their judgments compare to those of the entire panel (listen and interact), and then make a second round of judgments. Regardless of method, participating in a standard-setting workshop has been described as “a challenging judgmental process executed by well-meaning persons (with specific areas of relevant expertise and experience) who are subject to all the laws of social psychology and cognitive psychology (for good and for ill)” (J. Pellegrino, personal communication, February 25, 2019).

The importance of the human element in standard setting cannot be overstated. Human judgment is the cornerstone on which standard-setting deliberations are based and recommendations are formulated. Each standard-setting method is defined by its own approach to framing these judgments and guiding expert panelists to complete a specific cognitive-judgmental task. Human judgment is subject to influences and biases that play out as internal and external factors, and their potential impact on standard setting cannot be ignored.

In this chapter, we examine the cognitive-judgmental tasks involved in making standard-setting judgments. To explore this angle on standard setting, we draw on findings in judgment and decision-making research literature and consider what may be going on in the minds of standard-setting panelists when they participate in standard-setting workshops and make judgments about cut scores. We address three claims:

1. The cognitive-judgmental task (inherent in the standard-setting method) is a critical aspect of standard setting and should be central to the design, preparation, execution, and evaluation of the workshop.
2. In addition to the cognitive-judgmental task, there are other influences on panelist judgment and decision-making in standard setting. Some are design features and are desirable; some are not.

3. Standard-setting results should reflect a well-executed process that was led by experts who helped panelists make their judgments avoiding undesirable influences.

Following this overview, the second section of the chapter is a review of the history of standard setting, including the significance of major research and policy milestones. In the third section, we propose a *cognitive and social model of standard setting* that illustrates the cognitive and social influences on panelist judgment and decision-making in standard setting. At the center of this model is the cognitive-judgmental task required by a given standard-setting method. The fourth section provides a framework by which standard-setting methods can be categorized by the type of task required. The fifth section further explores these categories of methods by describing how panelists complete the task and the cognitive processes, heuristics, and biases that may influence how they do so.

The chapter then shifts, in the sixth section, to review how all of this is operationalized in the standard-setting workshop process and the related cognitive and social influences that can come from the workshop design and execution. In the seventh section, we provide a review of the research on standard setting that supports the claims made in this chapter and suggest directions for future research. The chapter concludes in the eighth section, with some predictions for the evolution of the practice in standard setting.

With this focus on the cognitive-judgmental task of standard setting, this chapter is not the typical introduction to standard-setting concepts or methods; there are excellent resources that provide this level of information for those new to this topic. Readers interested in this level of information are referred to the second section of this chapter, which outlines the publication of key resources in the history of standard setting, introducing specific methods and providing a comprehensive picture of the practices.

HISTORY OF STANDARD SETTING

The concept of setting standards has existed for decades, even before it was commonplace in educational and credentialing testing. Standards exist to help professionals identify acceptable or unacceptable values on measures in fields such as medicine (e.g., temperature of 100.4 °F or higher indicates a fever) and environmental safety (e.g., the maximum levels for more than 90 contaminants in drinking water; Environmental Protection Agency, n.d.).

Standard setting in educational and credentialing testing has a rich history. In Table 12.1, we present an overview of the major milestones in this history, including key publications, introduction of selected methods, and other events that illustrate how the practice of standard setting has developed since the 1950s.

As is evident in Table 12.1, standard-setting methods have proliferated in recent decades, workshop designs and procedures have evolved since the early 2000s, earlier debates about the defensibility of standard setting have quieted, and innovations have flourished.

Table 12.1 Selected Milestones in the History of Standard Setting

Milestones	Significance
1950s	
Lindquist (1951) <i>Educational Measurement</i>	The first edition of this volume provided a baseline on standard setting by including a discussion of measurement for classification and placement and predicting future achievement (e.g., Chauncey & Frederiksen, chap. 4; Cook, chap. 1).
Nedelsky (1954) <i>Absolute grading standards for objective tests</i>	Introduction of Nedelsky's method is "perhaps the first widely disseminated criterion-referenced method for setting cut scores" (Cizek & Bunch, 2007, p. 69).
1960s	
Glaser (1963) <i>Instructional technology and the measurement of learning outcomes: Some questions</i>	Another early reference to criterion-referenced measurement; criterion referencing established the need to differentiate levels of performance, that is, performance standards. See also Glaser and Klaus (1962).
1970s	
Angoff (1971) <i>Scales, norms, and equivalent scores</i>	Angoff introduced his idea for setting standards in a footnote in his chapter; numerous implementations since the original publication have adapted his idea and are captured under the umbrella of modified Angoff methods.
Shepard (1978) <i>Journal of Educational Measurement</i> [Special issue on standard setting]	Among other topics, Shepard highlighted the current debate on standard setting. Glass condemned judgmental standards as arbitrary and therefore unsafe; Linn argued that standards may be arbitrary but not necessarily capricious; Popham defended standard setting as an "important but largely unstudied problem" (p. 300); Levin condemned performance standards as an "appearance of accountability" (p. 316).
Other standard-setting methods introduced Ebel (1972) – Ebel method Berk (1979) – Contrasting Groups method Zieky and Livingston (1977) – Borderline Group method	
1980s	
Livingston and Zieky (1982) <i>Passing scores: A manual for setting standards of performance on educational and occupational tests</i>	The first practically oriented manual on standard-setting methods
Berk (1986) <i>A consumer's guide to setting performance standards on criterion-referenced tests</i>	Provided a comprehensive review of 38 standard-setting methods highlighting the range of approaches
Linn (1989) <i>Educational Measurement</i> (3rd ed.)	This third edition provided a brief discussion of standard setting in two chapters: Jaeger, chap. 14; Petersen et al., chap. 6

(continued)

Table 12.1 (continued)

Milestones	Significance
Other standard-setting methods introduced Jaeger (1982) – Structured Judgment process Hofstee (1983) – Hofstee Compromise method Beuk (1984) – Beuk Compromise method Legislation and professional standards AERA et al. (1985) <i>Standards for Educational and Psychological Testing</i> Public Law 100-297 (1988) establishes the National Assessment Governing Board to guide achievement-level setting for the National Assessment of Educational Progress (see Loomis & Bourque, 2001)	
1990s	
Shepard et al. (1993) <i>Setting performance standards for student achievement. A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels</i>	Evaluation of NAEP standard setting concludes that “the approach [to standard setting] is fundamentally flawed” (pp. 127–128).
Kane (1994) <i>Validating the performance standards associated with passing scores</i>	Kane defined a validity framework for supporting cut scores including procedural, internal, and external evidence to support the interpretation and use of performance standards.
Hambleton and Plake (1995) <i>Using an extended Angoff procedure to set standards on complex performance assessments</i>	Introduction to the extended Angoff method, which enabled standard setting for assessment activities scored in multiple levels
Lewis et al. (1996) <i>Standard setting: A bookmark approach</i>	Introduction of the Bookmark method, one of the most commonly used methods in educational achievement testing See also Lewis et al. (2012)
Impara and Plake (1997) <i>Standard setting: An alternative approach</i>	Introduction of the Angoff yes/no method, an alternative to modified Angoff, reflecting Angoff’s original ideas
Pellegrino et al. (1999). <i>Grading the nation’s report card: Evaluating NAEP and transforming the assessment of educational progress</i>	A second evaluation of NAEP standard setting, which concluded that “the current process for setting NAEP achievement levels is fundamentally flawed . . . the current processes are too cognitively complex” (p. 7)
Other standard-setting methods introduced Jaeger (1995a; 1995b)—Judgmental Policy Capturing method Plake et al. (1997)—Dominant Profile method Legislation and professional standards AERA et al. (1999)— <i>Standards for Educational and Psychological Testing</i>	
2000s	
Cizek (2001) <i>Setting performance standards: Concepts, methods, and perspectives</i>	First edited volume that is dedicated to standard-setting methods, approaches, and issues
Cizek (2005) <i>Applied Measurement in Education</i> [Special issue on vertically moderated standard setting]	Journal special issue focused on vertically moderated standard setting (also called “vertically articulated” standard setting) in response to reporting mandates from No Child Left Behind

Milestones	Significance
Buckendahl (2005) <i>Applied Measurement in Education</i> [Special issue on qualitative inquiries of participants' experiences with standard setting]	Special issue focused on participants' experiences in standard setting. The guest editor's introduction acknowledges the cognitive demands of standard-setting tasks and observes that the process involves translating "a participant's mental model of the target test taker . . . into judgments" (p. 219).
Cizek (2006) <i>Standard setting</i>	This chapter appeared in the first volume of the <i>Handbook of Test Development</i> , acknowledging the role of standard setting in overall test design and development.
Hambleton and Pitoniak (2006) <i>Setting performance standards</i>	Standard setting was addressed in its own chapter for the first time in <i>Educational Measurement</i> (4th ed.), providing a comprehensive treatment of existing methods and procedures.
Standard-setting methods introduced Jaeger and Mills (2001)—Integrated Judgment procedure Kingston, Kahl, Sweeney, and Bay (2001)—Body of Work method Plake and Hambleton (2001)—Analytical Judgment method Haertel (2002)—Briefing Book method Roeber (2002)—Reasoned Judgment method Sireci, Hambleton, and Pitoniak (2004)—Direct Consensus method Legislation and professional standards No Child Left Behind Act (2001) requires states to set academic achievement standards for state educational accountability tests NCCA (2004) <i>Standards for the Accreditation of Certification Programs</i> includes expectations for standard setting	
2010 to Present	
Kaftandjieva (2010) <i>Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six methods with an application to tests of reading in EFL</i>	Identified 62 distinct standard-setting methods (see Appendix 1); many were developed for language tests used in Europe and Asia and method names may not be recognizable to U.S. psychometricians (e.g., the basket procedure).
Cizek (2012a) <i>Setting performance standards: Foundations, methods, and innovations</i> (2nd ed.)	The second edition of the volume on standard setting contains multiple chapters that address standard-setting methods, approaches, and issues.
Phillips (2012) <i>The benchmark method of standard setting</i>	Introduction of an empirically based approach to standard setting, benchmarked standard setting. See also Ferrara et al. (2021)
Mattar et al. (2012) <i>Reviewing or revalidating performance standards on credentialing examinations</i>	Introduction to standards validation as an alternative to traditional standard setting, signaling evolution in standard-setting design and practice See also Landl (2019) for another treatment of standards validation.
Cizek and Earnest (2016) <i>Setting performance standards on tests</i>	This updated second edition of the <i>Handbook of Test Development</i> retained a separate chapter on standard setting, acknowledging the role of standard setting in overall test design and development.

(continued)

Table 12.1 (continued)

Milestones	Significance
Lewis and Cook (2020) <i>Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles</i>	Introduction to the embedded standard-setting approach, in which cut scores are established as part of the test development process; signals evolution in standard-setting design and practice
Reckase (2023) <i>The psychometrics of standard setting: Connecting policy and test scores</i>	This volume connected the psychometrics of standard setting to policy underlying the purpose of the specific standard setting.
Other standard-setting methods introduced	
Ferrara and Lewis (2012)—Item-Descriptor matching method	
Perie and Thurlow (2012)—Dominant Profile method	
Legislation and professional standards	
AERA et al. (2014) <i>Standards for Educational and Psychological Testing</i> includes three standards on standard setting (similar wording as in the 1999 edition)	
NCCA (2014) <i>Standards for the Accreditation of Certification Programs</i> includes expectations for standard setting	
Every Student Succeeds Act (2015) includes specific expectations for setting academic achievement standards (critical element 6.2)	
Domaleski et al. (2018), through CCSSO, creates <i>Establishing Performance Standards for School Accountability Systems</i>	

One driver of innovation in standard-setting methods has been the desire to improve the experience for panelists and ensure they can apply their expertise to make reasoned judgments following the guidelines within the method. Drawing on the extensive research in the field of standard setting and the authors' experience conducting these workshops, we recognize there are numerous influences on the judgmental process but no existing framework for organizing or modeling them. Therefore, we have developed a *cognitive and social model of standard setting*, which is detailed in the next section. We present that first because the remaining sections of the chapter discuss the aspects of this model and how the various influences (both desired and undesired) come into play during the workshop.

Historically, standard-setting methods have been classified as either test centered or test-taker centered to identify the focus of the judgmental task (Cizek, 2012c, p. 8; Jaeger, 1989). This has been a useful distinction because test-taker-centered methods require panelists to know the test takers and understand their knowledge, skills, and abilities and how they are taught, as well as the test content, whereas test-centered methods require panelists to bring their subject matter expertise about test content to the judgmental process. The research finding that panelists appear to understand test-taker-centered tasks better than test-centered tasks (Skorupski, 2012, p. 139) reinforces the distinction's usefulness. Hambleton and Pitoniak (2006, p. 440) recognized the need to expand this framework, given the increasing complexity of the design and scoring of educational achievement tests. They identified four categories, differentiated by the focus of panelist judgments: (a) item and scoring rubrics, (b) test takers, (c) test-taker work, and (d) score profiles. We agree with the distinc-

tions in both frameworks—that selecting a standard-setting method should take into consideration the types of judgments required of panelists and that the focus of these judgments should be aligned with the types of items on a test. We advance this thinking by reviewing cognitive-judgmental tasks required of panelists by various standard-setting methods.

To define standard-setting cognitive-judgmental tasks formally, we begin with a definition of *cognitive tasks*. According to Kester and Kirschner (2012),

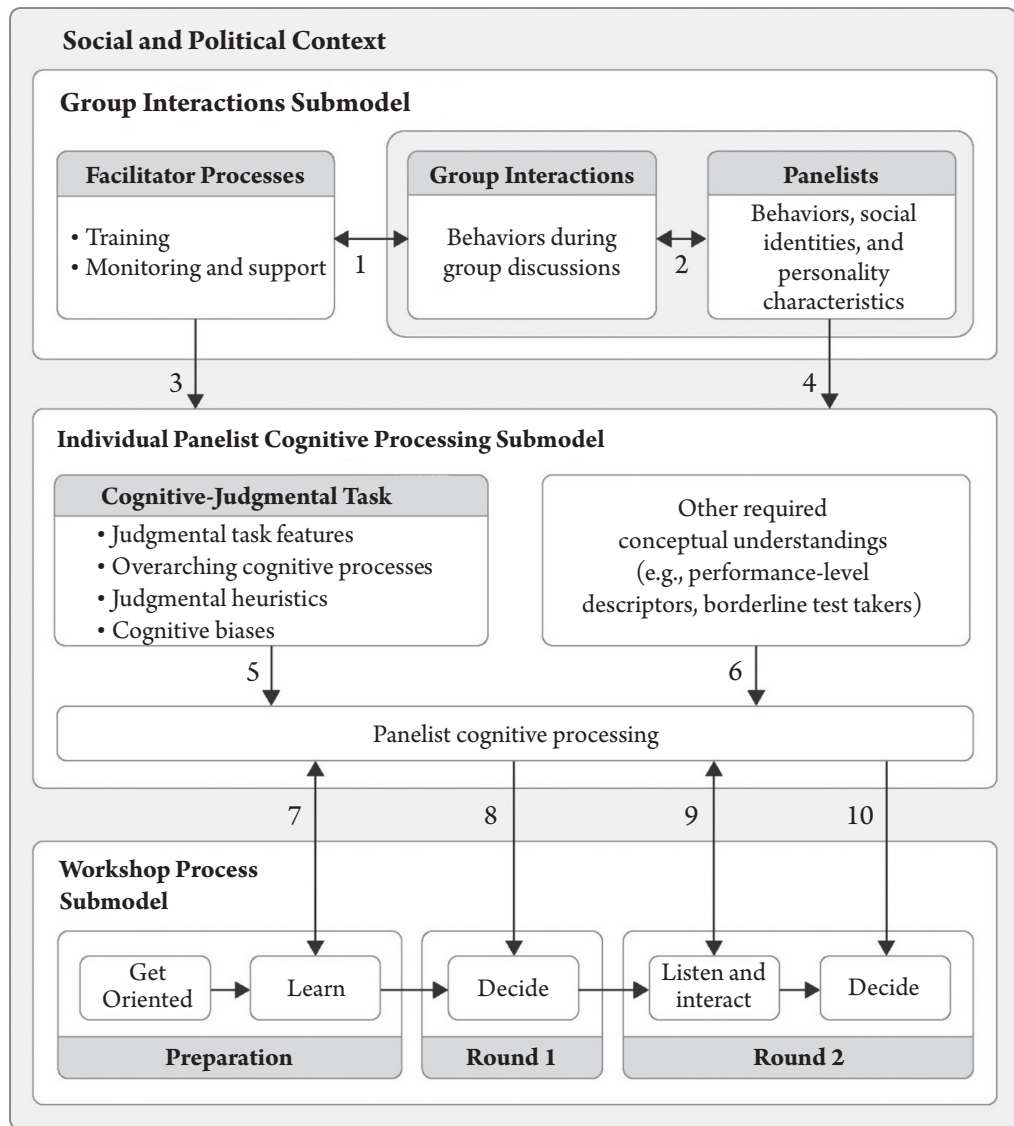
Cognitive tasks are those undertakings that require a person to mentally process new information (i.e., acquire and organize knowledge/learn) and allow them to recall, retrieve that information from memory and to use that information at a later time in the same or similar situation (i.e., transfer). (p. 619–620)

According to this definition, the components of cognitive tasks are (a) acquire and organize new learning and (b) retrieve and apply that learning to complete tasks that require cognitive processing. Standard-setting *cognitive-judgmental tasks* are the specified decisions that panelists are required to make to recommend cut scores. This term, cognitive-judgmental task, highlights the interrelatedness and interactions among the demands placed on panelists when they undertake a standard-setting method to recommend cut scores: (a) cognitive processes, (b) judgmental processes, and (c) a problem-solving task, in the form of a decision to be made (i.e., the recommendation required by the standard-setting method).

The “cognitive-judgmental tasks are the distinguishing features of the various standard setting methods” (Ferrara & Lewis, 2012, p. 255). For most methods, the specified judgment fully defines the cognitive-judgmental task (e.g., estimate a probability or percentage in the modified Angoff method) and provides information to calculate cut scores. For other methods, the specified judgment and its accompanying thinking process are a step toward making an explicit decision, in the form of directly recommending a cut score (e.g., in the Bookmark and dominant profile methods). Most methods require consideration of additional information to make a judgment (e.g., performance-level descriptors [PLDs], descriptions of borderline test takers¹—those who represent the key differentiators between levels). In all cases, the cognitive-judgmental task is intended to elicit a thinking process to arrive at a judgment (i.e., make a probability, matching, signal detection, or value judgment).

A COGNITIVE AND SOCIAL MODEL OF STANDARD SETTING

Our model of standard setting subsumes three submodels: a panelist cognitive processing submodel, a group interactions submodel, and a workshop process submodel (see Figure 12.2). At the center is the cognitive-judgmental task within the panelist cognitive processing submodel. The arrows within the model identify the influences among these components. Finally, the model reflects that standard-setting workshops and panelist judgments take place within a larger policy and sociocultural context.

**FIGURE 12.2****The Cognitive and Social Model of Standard Setting**

Note. The model illustrates facilitator processes, group interactions, panelist cognitive processing, and workshop processes that influence cut score recommendations. Boxes represent processes and elements in the three submodels; arrows represent hypothesized directions of influence.

- Arrow 1: Facilitator influence on group behaviors and interactions; group behaviors that require facilitator intervention
- Arrow 2: Group interactions that influence individual panelist thinking and behavior; individual panelist behaviors that influence group interactions. The gray box around “Group Interactions” and “Panelists” indicates that panelists behave and act as both individuals and group members.

- Arrow 3: Facilitator influence on individual panelist thinking and deciding (or cognitive processing and decision-making)
- Arrow 4: Panelist behaviors, identities, and characteristics that influence individual panelist thinking and decision-making
- Arrow 5: Influence of standard-setting task features, overarching cognitive processes, heuristics, and biases on panelist cognitive processing and decision-making
- Arrow 6: Influence of other required conceptual understandings on panelist cognitive processing and decision-making
- Arrow 7: Reciprocal influence of panelist cognitive processing while learning required concepts and procedures
- Arrow 8: Influence of panelist cognitive processing and the cognitive-judgmental task on Round 1 cut score judgments
- Arrow 9: Reciprocal influence of panelist cognitive processing during group discussions of judgments and Round 1 feedback
- Arrow 10: Influence of panelist cognitive processing and the cognitive-judgmental task on Round 2 cut score judgments

In this next section, we describe each submodel. We begin with the panelist submodel because panelist thinking and deciding are central to the standard-setting activities.

Individual Panelist Cognitive Processing Submodel

The *individual panelist cognitive processing submodel* accounts for the method's cognitive-judgmental task and related general cognitive processing (i.e., judgmental task features, overarching cognitive processes, judgmental heuristics, and cognitive biases), as well as other required conceptual understandings. A panelist's cognitive processing includes all of these aspects. Responding to requirements of the standard-setting cognitive-judgmental task is influenced by the other two components. This submodel is the focus of this chapter.

Group Interactions Submodel

The *group interactions submodel* accounts for the social actors who participate in standard-setting workshops and their interactions among these social actors. This submodel is supported by social psychology theoretical and empirical literature on group interactions and decision-making. Early on, Fitzpatrick (1989) discussed two processes by which panelists may change their minds following group discussion in a standard-setting workshop: normative influences based on social comparison (Festinger, 1954) and learning of new information in social contexts (Burnstein & Vinokur, 1975). In addition to these social influence processes, we propose explicit consideration of the individuals involved in a standard-setting workshop: the panelists and the facilitator.

One component of the group interactions submodel is the panelists themselves. Standard-setting panelists may be broadly representative of stakeholders of the

assessment program, which implies some degree of diversity. In addition to this intentional diversity, panelists also bring varied personality characteristics, personal attributes, and points of view that may not be known during workshop planning. A second component of the group interactions submodel poses the influence of the panelist group on the standard-setting process. Panelist behavior and interactions are defined in small working groups, often organized as separate tables, rather than as coexisting individuals, and observable and inferable behaviors must be considered to understand their processes fully (Bonito & Sanders, 2011). The final component of the group interactions submodel is the workshop facilitator, who guides panelists through every step of the workshop as detailed in the “Workshop Designs, Processes, and Practices” section of this chapter. Because of space limitations, the group interactions submodel is not covered in detail in this chapter.

Workshop Process Submodel

The *workshop process submodel* describes three phases and five subsumed steps that are discernible in the agendas for most standard-setting workshops and methods, though the emphasis and amount of time allotted to any step can vary considerably. The three phases include preparation, Round 1 judgments, and Round 2 judgments. (Some workshops include more than two rounds.) Standard-setting workshop procedures generally follow similar steps and contain similar content and are customized for a specific method and workshop (see detail in the section “Workshop Designs, Processes, and Practices” later in this chapter).

Social and Political Context

These submodels (and the activities therein) exist within the larger social and political context, which is specific to each testing program. In this chapter, we focus on standard-setting practices within educational testing and credentialing testing. Although there are similarities in the design, execution, and evaluation of these activities, the social and political context is unique.

Educational tests are used to evaluate student achievement against performance standards as articulated through PLDs (see Huff et al., this volume). This may require setting a single cut score (e.g., pass or fail on a high school end-of-course examination) or multiple cut scores (e.g., to distinguish levels of performance). The results from these educational tests are used primarily for accountability purposes (as required under federal law); to evaluate school performance and improvement needs; to identify student groups that may require additional, specialized attention; and to measure achievement growth. Standard setting in education, therefore, represents high-stakes decisions for schools, districts, states, and, in some cases, students. These observations shared in this chapter with respect to educational testing are also relevant to setting standards for alternate assessments for students with significant cognitive disabilities (e.g., Elliott & Kettler, 2015) and for assessments of English language proficiency (e.g., Chalhoub-Deville & Deville, 2006). Both types of educational tests

are required for public accountability reporting, specifically for these unique student groups. Panelists for setting standards for these tests are expert and experienced in both the content standards and the unique learning needs and performance capabilities of these groups.

Credentialing testing covers examinations that are used to determine eligibility for licensure, certification, and certificates. The purpose of a passing standard typically is to determine whether a candidate (i.e., a test taker) has met a minimum standard for safe and effective practice within the profession. For that reason, most credentialing examinations result in either a “pass” or “fail” determination. The context of the standard setting, therefore, includes high-stakes decisions where the ultimate result will have implications for individual candidates’ current and future professional careers as well as the safety and security of the public. Additional stakeholders include employers and training programs. The practices (e.g., standard setting) of many credentialing programs are often reviewed by third-party accreditation organizations against standards demarcating acceptable practice.

Relationships Among the Submodels and Components

The 10 arrows shown in Figure 12.2 indicate influences between components in the three submodels. Some influences are unidirectional, as in the influence of facilitator processes on individual panelist cognitive processing (e.g., Arrow 3). Others are bidirectional, for example, the reciprocal adjustments between workshop facilitators and group behavior (Arrow 1).

Cognitive processing operates within each panelist and is not directly observable during a workshop. Panelist processing and decision-making are enacted in the three phases and five steps in the workshop process submodel. In the Learn step, influences on the panelist submodel function in both directions: As panelists learn about the cognitive-judgmental task and other required conceptual understandings, their cognitive processing develops, becomes more efficient, and, hopefully, conforms to the method’s intended cognitive-judgmental task. A similar developmental process holds true as panelists process insights and develop shared understandings during discussion of feedback, in the Listen and Interact step. Panelist processing and deciding also connects to the Decide steps. We hypothesize that this connection and its influence operate in one direction primarily, though there seems little doubt that many panelists continue to process their decisions after they have made them. In the meantime, group interactions are likely to influence individual panelist processing and decision-making, both while group interactions are occurring (e.g., discussions of feedback between rounds) and during subsequent independent thinking and deciding.

We have proposed this cognitive and social model of standard setting using theory, concepts, and empirical findings in cognitive and social psychology and standard-setting literatures. We now turn our attention to the center of the model and the central

focus in standard setting: panelists and their cognitive processing. To do so, we draw on the judgment and decision-making branch of cognitive and social psychology, which has been influential in economic decision-making and business management and perhaps is most closely associated with the work of Kahneman (e.g., 2011) and colleagues. The process of standard setting involves humans who are trained to use specific types of information to make decisions in order to recommend cut scores. However, human decision-making is susceptible to judgmental errors and biases. As we will show, the cognitive-judgmental tasks that define the most widely used standard-setting methods are functionally different and are susceptible in different ways to decision errors and biases. We begin with an overview of cognition in decision-making before defining and contrasting the cognitive-judgmental tasks for eight well-documented standard-setting methods. Later, we consider other factors that influence standard-setting decisions under different methods.

TAXONOMY OF COGNITIVE-JUDGMENTAL TASKS FOR STANDARD-SETTING METHODS

With that overview as background, we now introduce a taxonomy of cognitive-judgmental tasks for standard-setting methods that can help practitioners be clear about the cognitive-judgmental task they are asking panelists to undertake. We propose this *taxonomy of cognitive-judgmental tasks* to illustrate the similarities and differences among standard-setting methods in terms of the cognitive challenges they pose to panelists and provide examples from frequently used methods (see Table 12.2). In accordance with the focus of this chapter, we will not fully describe each of these methods. For additional detail on these (and other methods), the reader is referred to the original research on each method, as well as texts that cover multiple methods (see citations in Tables 12.1 and 12.2).

The hypothesized cognitive processes involved are determined and influenced by the features of the method's standard-setting task and the cognitive and other demands on panelists that we illustrate in this model. The cognitive-judgmental task defines each standard-setting method and is central to the cognitive-social model. The model illustrates the many cognitive and social influences on decision-making in a standard-setting workshop.

We have organized the taxonomy around four types of cognitive-judgmental tasks by drawing on findings in cognitive psychology and judgment and decision-making empirical literatures. These are:

1. *Probability judgments*: State a probability (directly or indirectly) or estimate a percentage as part of the process of recommending cut scores.
2. *Matching judgments*: Match test items, test-taker work samples, test-taker score profiles, or test takers themselves with achievement-level descriptors or descriptions of borderline test takers.

Table 12.2 Taxonomy of Cognitive-Judgmental Tasks With Key Elements of Eight Standard-Setting Methods (and Two Variations), Including Paraphrased Statements of Cognitive-Judgmental Tasks

(1) Probability judgments			
Angoff, modified			
Cognitive-judgmental task: Estimate the (a) probability that borderline test takers will respond successfully to each item or (b) percentage of borderline test takers who will respond successfully (Angoff, 1971; Plake & Cizek, 2012). Typical uses: Most widely used in credentialing testing; moderate use in K–12 educational testing			
Item formats: Dichotomously scored items	Data requirements: Item p values and impact data are not required, but panelists would review them if provided	Other required understanding: Test items Descriptions of borderline test takers	Cut score recommendations: Sum of item-level judgments ^a ; calculated by panelists or a psychometrician
Angoff yes/no			
Cognitive-judgmental task: Indicate which items you expect a borderline test taker to answer correctly—yes or no—where the expectation is a probability judgment or perceived likelihood (Angoff, 1971, p. 514; Impara & Plake, 1997). Typical uses: Most widely used in credentialing testing; moderate use in K–12 educational testing			
Item formats: Dichotomously and polytomously scored items	Data requirements: Item and impact data are not required, but panelists would review them if provided	Other required understanding: Test items and polytomous item score levels ^b Descriptions of borderline test takers	Cut score recommendations: Sum of item-level judgments; calculated by panelists or a psychometrician
Bookmark			
Cognitive-judgmental task: Place a bookmark at the first location in the ordered item booklet where borderline test takers have demonstrated achievement sufficient to be considered <i>Proficient</i> (<i>Basic, Advanced</i> ; Lewis & Cook, 2020). Alternative statement: The bookmark represents a judgment of the divide between items that a student at the threshold of a performance level (the minimally proficient student) should be able to answer correctly from those items that are not necessary to demonstrate mastery of at that performance level. Typical uses: Most widely used method in K–12 educational testing			
Item formats: Dichotomously and polytomously scored items	Data requirements: Dichotomous item and polytomous score level difficulty locations on an IRT scale, using a response probability criterion, ordered by difficulty	Other required understanding: Scale locations of dichotomous and polytomous items PLDs Response probability criterion	Cut score recommendations: Panelist judgment (bookmark placement); translated to score scale by a psychometrician

(continued)

Table 12.2 (continued)

Other probability-based methods The direct consensus (Sireci et al., 2004) and Nedelsky (Nedelsky, 1954; Zieky, 2001, p. 21) methods are included in the probability judgment category.			
(2) Matching judgments			
(a) Match items or test-taker work samples to PLDs			
Body of work			
Cognitive-judgmental task: Match test-taker work samples to PLDs (Kingston et al., 2001; Kingston & Tiemann, 2012). Typical uses: Most widely used in K–12 educational testing for performance and portfolio assessments			
Item formats: Polytomously scored test-taker work samples, including performance tasks and essay prompts, typically contained in a portfolio	Data requirements: All test-taker responses are scored and sorted from lowest to highest total scores (or vice versa); scores are not revealed to panelists	Other required understanding: Test-taker work samples, ordered by overall score PLDs	Cut score recommendations: Cut scores are the difference between average scores of work samples that are matched to each performance level; calculated by a psychometrician
Item-Descriptor Matching			
Cognitive-judgmental task: Step 1. Match the content knowledge, skills, and cognitive processes that each item requires (i.e., response demands) to the knowledge and skill capabilities in PLDs. Step 2. Find the location in the threshold region between two adjacent levels that best distinguishes the levels, this time by matching threshold region items to adjacent PLDs (Ferrara et al., 2002; Ferrara & Lewis, 2012). Typical uses: Used most widely in K–12 educational testing, sometimes with alternate assessments			
Item formats: Dichotomously and polytomously scored items	Data requirements: Item and polytomous score level difficulty locations on an IRT scale, using a response probability criterion, ordered by difficulty	Other required understanding: Scale locations of dichotomous and polytomous test items, ordered by difficulty PLDs	Cut score recommendations: Panelist judgment (location); translated to score scale by a psychometrician
(b) Match score profiles to PLDs			
Dominant profile			
Cognitive-judgmental task: Match score profiles to descriptors, then state their conjunctive and compensatory matching rules; for example, <i>Require a score of 3 on two tasks and at least a 2 on the other three tasks</i> (Cizek & Bunch, 2007; Perie & Thurlow, 2012, pp. 366–388; Plake et al., 1997). Typical uses: Field tested for National Board for Professional Teaching Standards certification assessments; used in credentialing testing sometimes as an adaptation of judgmental policy capturing (C. Buckendahl, personal communication, May 22, 2020)			

(c) Match test takers to performance level groups			
Borderline group			
Cognitive-judgmental task: Categorize individual test takers into clearly competent, clearly inadequately prepared, and on the cusp between competent and not competent status, based on PLDs (Cizek & Bunch, 2007, p. 113; Livingston & Zieky, 1982). Typical uses: Used for the Smarter Balanced K–12 educational testing program (Smarter Balanced 2015; 2017)			
Item formats: Test items are not examined, though test content is considered	Data requirements: Test scores are required to compute the cut score; panelists are unaware of test-taker scores	Other required understanding: Panelist judgments of passing/ borderline/ failing status of known test takers Descriptions of passing and failing test takers are not specified	Cut score recommendations: Average score of test takers classified as borderline; calculated by a psychometrician
Contrasting groups			
Cognitive-judgmental task: Categorize individual test takers into mastery/nonmastery or pass/fail status (Berk, 1976; Cizek & Bunch, 2007, p. 107). Typical uses: Used frequently in employment testing (C. Buckendahl, personal communication, May 22, 2020)			
Other matching-based methods The analytical judgment (Cizek & Bunch, 2007; Plake & Hambleton, 2001; Zieky et al., 2008) and integrated judgment procedure (Jaeger & Mills, 2001) methods are included in the matching category.			
(3) Threshold-signal detection judgments			
The performance profile (Morgan, 2004; Morgan & Perie, 2004, Table 2; Perie & Thurlow, 2012) and reasoned judgment (Roeber, 2002) methods are included in the threshold-signal detection category.			
(4) Value judgments			
Beuk (Beuk, 1984; Thompson, 2022), briefing book (Haertel, 2002; Haertel et al., 2012), Ebel (1972), Hofstee (Hofstee, 1983), judgmental policy capturing (Jaeger, 1995a; 1995b), and the structured judgment process (Jaeger, 1982) methods are included in the value judgment category.			

Note. IRT = item response theory; PLD = performance-level descriptor.

^aWhen panelists understand how cut scores are calculated, they can adjust item-level judgments to achieve a desired test cut score.

^bScore levels may be expressed as a rubric score (e.g., 1, 2, 3, 4) or rubric score-level location or threshold location on an IRT score scale.

3. *Threshold-signal detection judgments*: Discern when a score profile's indication of the requirements of one performance level shifts to the requirements of the adjacent performance level.
4. *Value and other judgments*: Make value and other judgments to recommend cut scores.

Probability Judgments

Many of the decisions we make in our lives are based on our beliefs about the likelihood of events and other outcomes. We express our beliefs using phrases like “I think that . . .,” “Chances are that . . .,” and so forth. We may or may not be completely

aware, when we make predictions and attach our judgment of the likelihood that the prediction will prove true, that we are making probability judgments. In standard setting, panelists are required to express these “subjective probabilities” (Tversky & Kahneman, 1974; see also Kahneman, 2011, p. 419) in the modified Angoff, Bookmark, direct consensus, and Nedelsky methods. Tversky and Kahneman (1974) and innumerable subsequent studies and research syntheses (e.g., Nickerson, 2004; Plous, 1993) concluded that “people rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations . . . [that can] lead to severe and systematic errors” (see Kahneman, 2011, p. 419).

Probability judgments are prone to judgmental errors that arise from cognitive biases. For example, people ignore base rates² (Kahneman, 2011, p. 88) when they judge the likelihood of a future event, they do not evaluate the quality of evidence they use to make these judgments, and they use representativeness³ (e.g., stereotyping; Kahneman, 2011, pp. 149–153) to make these judgments. Research also has shown that some situations tend to elicit nonprobabilistic mindsets (e.g., Rottenstreich & Kivetz, 2006) where people do not judge the likelihood of an event (i.e., its probability) and, instead, use intuitively generated, nonprobabilistic rules, heuristics, or rationales. For example, cognitive biases like “what you see is all there is” (WYSIATI⁴; Kahneman, 2011, pp. 85–88) and our need for coherence (Kahneman, 2011, pp. 75–76) can lead us to believe the explanatory stories we construct about the likelihood of a future event rather than evaluate the quality and relevance of available evidence as we make probabilistic and other judgments.

Panelists probably construct an internal representation of the knowledge and skills possessed by borderline test takers, based on a description that adds to the cognitive load of the probability judgment and, likely, the complexity of making probability judgments. This may lead panelists to use the substitution heuristic⁵ (e.g., Kahneman, 2011, p. 97) to reduce load and complexity. In one study, Baron and Kannan (2013) found that panelists may estimate high probabilities of correct responses for borderline test takers when they see the knowledge and skills measured by an item listed explicitly in the borderline test-taker description. In contrast, they may estimate a lower probability of a correct response for borderline test takers when they encounter items that measure knowledge and skills not explicitly listed in the borderline test-taker description.

Katz (2019) offered a framework of standard-setting panelist cognition for making probability judgments, specifically for the modified Angoff method. He built his framework on others’ studies of how people estimate numerical values (e.g., likely future stock values), preferences (e.g., which is the better of two products), and probabilities (e.g., the likelihood that someone is infected with COVID-19) using attributes of the situation and levels of the attributes. He illustrated the role that specific heuristics and biases can play in modified Angoff estimates. The framework acknowledges that people ignore information cues, place too much focus on some cues, and take shortcuts when they estimate.

Matching Judgments

People match focus entities (e.g., a salmon, a test item) to target categories (e.g., fish, easy items) using subjective judgments about the degree to which the focus entity is representative of an entity group and the features of the target category (Murphy, 2004, p. 373; Wills, 2013, p. 352). People use a variety of matching strategies, including “category-based induction,” to categorize and “make predictions about novel items” (Murphy, 2004, p. 243) and their category membership. Working in the opposite direction, people may use “property induction” (Hayes & Heit, 2013, p. 618) to infer properties from a category, such as knowledge and skill requirements in PLDs, to enable matching of test items and their response demands.

We hypothesize three subcategories of matching judgments in standard setting:

- Match items or test-taker work samples to PLDs or descriptions of borderline test takers (e.g., Body of Work or ID Matching methods)
- Match score profiles to PLDs or descriptions of borderline test takers (e.g., dominant profile method)
- Match test takers to PLDs (e.g., borderline method)

All three subcategories entail matching entities (i.e., items, work samples, and score profiles) to other entities (i.e., PLDs, descriptions of borderline test takers). Formal models of categorization may involve multiple competing processes, which have been shown to include probability matching (Wills, 2013, p. 352). While these three matching subcategories may appear to be unique, we hypothesize that cognitive processing is highly similar for all three and that matching focus entities and target entities is likely to require activation of similar declarative knowledge networks.

Threshold-Signal Detection Judgments

The concepts in signal detection theory reach back at least to the psychophysics research of Fechner in the late 1800s (Briggs, 2022, chap. 2; Swets, 1996, p. 7). Electrical and radar engineers developed the receiver (or relative) operating curve method in the 1940s to evaluate peoples’ diverse sensitivities to target stimuli (e.g., enemies on a battlefield). Detecting a signal amid noise is a main function of conscious attention. In psychology, the focus has been on perceptual detection of stimuli. We hypothesize that signal detection is the primary cognitive-judgmental task in the two methods in this section of the taxonomy (performance profile and reasoned judgment) and the secondary cognitive function in other methods (Bookmark, ID Matching, and dominant profile).

In standard setting, the signal is the detectable difference between two adjacent levels of performance or test-taker groups, indicated in items, score profiles, and scoring rubrics; the noise is competing and incomplete information, which obscures the information in the signal that is most useful in detecting a threshold. In detecting signals, panelists also consider, for example, the consequences associated with false-positive and -negative decisions, as in “If I recommend this cut score, then X percentage of test takers will fail.”

Nichols et al. (2010) proposed standard setting as a measurement process in which “the construct being measured is the panelists’ representation of student performance at the threshold of an achievement level” (p. 14). They hypothesize that panelists construct internal representations of a scale and a value on that scale (both are concepts in statistical signal detection theory) that indicates where the signal has changed from one category to another (e.g., *Nearing Proficiency to Proficient*, nonqualified to qualified). The predominance of signal detection in Round 1 for some standard-setting methods is likely reduced in subsequent rounds as a result of feedback and discussion.

Several of the standard-setting cognitive-judgmental tasks in other categories in the taxonomy contain an element of signal detection. For example, statements of the cognitive-judgmental task for the Bookmark method make the role of threshold-signal detection clear. In fact, it is not unreasonable to hypothesize that signal detection is part of all methods in the probability category of the taxonomy because all require making judgments about borderline test takers or the minimum threshold between adjacent performance levels.

Value Judgments

Value judgments involve “assigning a value (such as good or bad) to something” (Merriam-Webster, n.d.). *Psychology Wiki* elaborates: “A value judgment is a judgment of the rightness or wrongness of something, based on a particular set of values or on a particular value system. Value judgments are statements of subjectivity” (Psychology Wiki, n.d.). Some methods in this section of the taxonomy require percentage estimates and matching judgments. However, these judgments are made within a framework of value judgments.

Standard setting involves making judgments that are inherently value influenced, as in “Which cut score is the best or right score for these test takers and this test?” As the *Standards* state, “cut scores embody value judgments as well as technical and empirical considerations” (AERA et al., 2014, p. 101). However, methods in this category task panelists with making holistic cut score judgments based on the overall assessment and/or expected performance of test takers.

The proposed taxonomy is further exemplified in Table 12.2, which outlines how some common standard-setting methods serve as examples of each category. In addition to providing a description of the cognitive-judgmental task for each method, the table includes descriptions of how the method is commonly used, the item formats for which it is appropriate, the data required to implement the method as described, additional information panelists must understand to complete the cognitive-judgmental task, and how the cut score recommendations are derived from the judgments. Because of space limitations, the additional details are provided for those methods that are based on probability and matching judgments since these are most commonly used in current practice.

FACTORS INFLUENCING COGNITIVE PROCESSING IN STANDARD SETTING

Some psychological theories structure human decision-making as if people decide rationally because they seek optimal outcomes. A thoughtful look around the world suggests that this is not the case. The empirical and theoretical literature on judgment and decision-making (e.g., Kahneman, 2011; Kahneman et al., 1982; Plous, 1993) and business management textbooks based on this literature (e.g., Bazerman & Moore, 2013; Hastie & Dawes, 2010) emphasize the role of judgmental heuristics and cognitive biases in human thinking and deciding. More specifically, they illustrate that human thinking, judgment, evaluations, and decisions are strongly influenced by (a) heuristics that provide shortcuts for difficult (and not so difficult) thinking tasks, (b) biases that often are not part of our conscious awareness or recognized in executive functioning, and (c) other cognitive processes and factors. Some of these heuristics, biases, and other cognitive processes are applicable as panelists engage in the standard-setting cognitive judgmental tasks.

Simon's bounded rationality (Simon, 1956; Speekenbrink & Shanks, 2013, p. 688), where people satisfice by finding good-enough, rather than optimal, solutions, seems consistent with the characteristics of standard-setting cognitive-judgmental tasks. For example, panelists estimate probabilities and percentages, place bookmarks based on the probability of a correct response, align items or collections of evidence with PLDs, and so forth. While we have observed panelists working diligently to make reasoned decisions in standard-setting workshops, we also have observed them accepting the practical reality of making satisficing decisions. For example, panelists usually accept and are willing to work with empirical item difficulties even after estimating a different item difficulty (see Clauser, Mee, et al., 2009, 2013).

In standard setting, panelists follow the cognitive-judgmental task to make decisions about what cut score(s) to recommend. However, panelist cognitive processing is not solely influenced by the method's cognitive-judgmental task. Figure 12.2 shows how various factors, such as judgmental task features, overarching cognitive processes, judgmental heuristics, and cognitive biases that influence panelist processing and decision-making, interact with the demands of a method's cognitive-judgmental task. These factors, which can undermine the intent and valid implementation of a prescribed judgmental task, are described in this section.

Specifically, in this section of the chapter, we offer hypotheses about the influences of these factors on two of the categories within the taxonomy: probability judgments (i.e., Angoff and Bookmark methods) and matching judgments (i.e., Body of Work, ID Matching, Borderline Group, Contrasting Groups, and the Dominant Profile methods). These two categories include the most widely used standard-setting methods up to this time.

Our goal in proposing these hypotheses is to identify the heuristics, biases, and other cognitive factors that operate on decision-making in these methods and, therefore,

should be addressed in workshop design and panelist training. You will notice at least three things about these hypotheses. First, in a small number of instances where relevant research findings are available, we report the findings rather than generate new hypotheses. Second, our hypotheses indicate that, often, different standard-setting methods may be susceptible in different ways to influences of these factors. In other cases, a hypothesized influence may be similar across methods. While a reader may or may not find these hypotheses plausible, our goal is to stimulate thinking about judgment and decision-making in standard setting and susceptibilities of methods to specific types of judgmental errors. In a later section of this chapter, we provide considerations for panelist training and workshop facilitator monitoring in mitigating these undesirable influences on panelist judgments.

We now discuss the various factors that influence panelist cognitive processing in standard setting organized under the sections of judgmental task features, overarching cognitive processes, judgmental heuristics, and cognitive biases.

Judgmental Task Features

These are characteristics of the cognitive-judgmental task, such as cognitive load, cognitive strain, and cognitive complexity, that influence the cognitive challenges of the task.

Cognitive load and cognitive strain describe the demand on working memory or the level of effort required for a task. Such demands reduce memory capacity for new learning and decisions (see Kahneman, 2011, pp. 59–60, on cognitive strain; see Sweller, 1988, on cognitive load). Therefore, such factors may have an influence on panelist cognitive processing in probability and matching judgment methods where the panelists follow the cognitive-judgmental task by evaluating the knowledge and skill response demands for each item or piece of work and making the appropriate judgment. They must do this while keeping in working memory the description of the borderline test taker (or the expectations in the PLDs as in the Bookmark method) as well as their judgments about previous items to manage consistency. We hypothesize that this level of cognitive load (or strain) is likely to be higher in Round 1 as compared to subsequent rounds of judgments.

Cognitive complexity refers to the number and complexity of mental interactions required to make a decision (Bieri, 1955). Therefore, the cognitive complexity of a standard-setting judgmental task describes the number of factors that must be considered when making a judgment. In probability judgments, the cognitive complexity comes from having to integrate item knowledge and skill demands with features of the borderline test-taker description while maintaining consistency with previous judgments (i.e., tell a coherent story; see Kahneman, 2011, p. 87). In matching methods, complexity arises from having to identify item knowledge and skill response demands (also a challenge in probability methods). They then must align those demands with knowledge and skill expectations in PLDs or descriptions of borderline test takers, which defines minimal mastery of a content and skills domain. ID matching appears to require simply matching item response demands and descriptors, but no research has examined this

claim of simplicity yet. Empirical research illustrates that people use different strategies for matching entities. Integrating item and descriptor features is shown to be complex (see Katz, 2019, for a review of panelist cognition models). Cognitive complexity may be more challenging for the Body of Work method, where panelists must integrate their holistic judgments about the qualities of one or more pieces of evidence, and the dominant profile method, where panelists must analyze score profiles, integrate differences among similar profiles, and match those to descriptors. The complexity of matching known people (e.g., students, credential candidates) to a performance level may be the least cognitively complex matching task because panelists probably imagine one or more known people (or borderline test takers) with little specification of the basis for making those judgments. Research indicates that panelists seem to understand test-taker-centered tasks better than item-centered tasks (Skorupski, 2012, p. 139).

Some methods may lead to a propensity for algorithmic thinking where panelists develop mental rules to increase cognitive ease when making repetitive judgments. Specifically, panelists may begin the judgmental process with systematic, evidence-based analytic judgments (see Kahneman, 2011) but then progress toward more “automated” judgments to reduce cognitive load and strain. We hypothesize that the propensity for algorithmic thinking is increased when judgments are required for a large number of items (e.g., Angoff, Bookmark). An example of this is when panelists may make Angoff judgments quickly based on use item difficulty values (e.g., p values) without fully following the cognitive-judgmental task. Further, panelists may make automatic judgments for items that appear similar, influenced by the representativeness heuristic and halo effect (both of which are described later in this section), and may confirm snap judgments about items with little reflection. In matching methods, we have observed that panelists match sequences of items, one sequence after the other, rather than item by item, in an ordered item book (OIB) until they detect a change in ID matches. Panelists may use similar algorithms to reduce cognitive load (not to mention boredom) for other matching methods.

Overarching Cognitive Processes

Humans use processes such as framing effects and continuous monitoring to manage various types of cognitive tasks in our daily lives, but they can greatly impact our thought process. In standard setting, we hypothesize that panelists sometimes apply overarching cognitive processes such as framing effects to the standard-setting tasks.

Framing effects are seemingly inconsequential variations in a problem statement (e.g., how a cognitive-judgmental task is stated) that can lead to significant differences in decisions (Kahneman, 2011, p. 272; Tversky & Kahneman, 1981). Cognitive-judgmental task statements for probability methods such as “Would a borderline test taker respond correctly to each item on a test?” and “Place a bookmark at the first location in the OIB where borderline test takers have demonstrated achievement sufficient to be considered *Proficient* (etc.)” are stated unambiguously and likely would be interpreted in similar ways by most panelists. Cognitive-judgmental task statements for matching

methods should be familiar and even routine for teachers. For example, “Which PLD most closely matches the knowledge and skill response demands of this item?” (from ID Matching) is part of what teachers and professional preparation program trainers do when they break down the components of learning tasks for students as part of instructional planning. Still, panelists are likely to use some heuristics to reduce the difficulty of making the stated matching tasks, as indicated in two studies (Hein & Skaggs, 2009, 2010). When standard-setting facilitators state the cognitive-judgmental task unambiguously and repeatedly, the chance that panelists will succumb to framing effects and modify the problem statement is reduced.

Continuous monitoring and basic assessment process is a cognitive processing strategy where we pay automated, unconscious attention to aspects of a task to overcome dissonance and recover comfort (i.e., using automatic, quick, and easy decisions; Kahneman, 2011, p. 20). Standard-setting panelists are likely to feel some dissonance and discomfort as they make probability judgments for some items. They may feel uncomfortable when they believe that some items in an OIB are misordered, or when they judge item difficulties against what they understand about borderline test takers, or when they compare their judgments to empirical item difficulty, as is often practiced in Angoff method workshops. In matching methods, panelists may specifically monitor their judgments about the elements in the entities that they are required to match. They may monitor the consistency of current judgments with past judgments (i.e., to construct a coherent story; see Kahneman, 2011, p. 87). ID Matching workshop facilitators have observed that panelists report discomfort when they feel an item is “out of order”—that is, not in an expected location in an OIB. Some panelists may not be comfortable with matching items in the threshold region (i.e., items in the OIB that they perceive are located between two performance levels).

Judgmental Heuristics

Heuristics are mental shortcuts humans take to simplify a judgmental process. We have observed panelists use several heuristics such as substitution, availability, representativeness, and recognition as they engage in the cognitive-judgmental tasks required by the standard-setting method.

Substitution is used when we replace a challenging judgmental task with a simpler heuristic question and then map the judgment about the simpler question back to the target question (Kahneman, 2011, p. 97). In probability judgments, for example, panelists may substitute the complex judgment—“Can Grade 6 borderline test takers be expected to solve for x in $2x = 5$?”—with a simpler judgment such as “Can Grade 6 borderline test takers be expected to solve for x in simple equations like this one?” In matching methods, panelists may use a variety of substitution judgments when matching entities becomes particularly difficult. For example, they may find ways to simplify their judgments about matching diverse pieces of evidence to PLDs in the Body of Work and dominant profile methods. In

ID Matching, panelists may designate items to the threshold region between adjacent performance levels simply to reduce the cognitive strain of making a definite ID match, with the intention to make more definitive decisions later in the judgmental process.

When working with scales and values, we are prone to intensity matching and values translation. This may occur when panelists judge the intensity or amount on an underlying scale (e.g., a probability or percentage, degree of match) and translate it to standard-setting judgments (Kahneman, 2011, p. 94). In probability methods, panelists may calibrate their judgments about borderline test takers by estimating, for example, that borderline test takers would be able to respond successfully to an item sometimes, which might suggest a lower probability, or most of the time, which might indicate a higher probability. In matching methods, panelists may use an internal scale of the closeness of matches between entities to reduce cognitive strain and then translate those judgments to the specific requirements of the Body of Work, ID Matching, and other matching methods. In addition, panelists may rely on threshold-signal detection first and then match entities rather than implement the specified cognitive-judgmental matching task.

We use the availability heuristic when we judge the probability or frequency of a class or event based on the ease of bringing instances and occurrences to mind and disregarding other evidence (Tversky & Kahneman, 1974). In probability methods, panelists may make judgments by calling to mind a test taker they know rather than based on the borderline test-taker description. This hypothesis also represents the substitution heuristic. In matching methods, panelists may match collections of evidence, items, score profiles, and known people to descriptors because they can recall such instances easily, rather than a deeper, more detailed consideration.

We use the representativeness and recognition heuristics when we judge the probability that an entity matches another entity or belongs in the same group as other entities, based on how much they resemble recalled prototypes or other members of the group, while disregarding other crucial judgmental information, as in stereotyping people (see Tversky & Kahneman, 1974). Recognition is a faster process based on instantaneously perceiving an entity as an example or not an example of a group (see Gigerenzer & Goldstein, 1996). In probability methods, panelists may make judgments for some items that are similar to other items in terms of knowledge and skill response demands. (This is an illustration of a positive heuristic effect that we might call efficiency, unless panelist recognition judgments are inaccurate.) Panelists may give more importance and expect all test takers to respond correctly to items that they recognize in their own area of specialization (e.g., literature, geometry) more often than items from other specialization areas. In matching methods, panelists may match collections of evidence, items, score profiles, and known people to descriptors as an approximation, as in “This match is probably close enough.” Research suggests that we can expect panelists to use a variety of matching strategies (Kahneman et al., 2021) unless the facilitator specifies the matching process.

Cognitive Biases

Finally, cognitive biases can be errors in human thinking that can impact how we interpret information and make decisions. We hypothesize that a number of cognitive biases, such as confirmation bias and anchoring and adjustment bias, can impact how panelists engage in the cognitive-judgmental tasks required by the standard-setting method. Some of these cognitive biases are discussed here.

Overconfidence bias occurs when we have high confidence about the accuracy of inaccurate judgments. The impact of this bias is most extreme in tasks of great difficulty (Lichtenstein et al., 1982, pp. 314–315). Further, people typically have great confidence in judgments that they base on heuristics (Slovic et al., 1982, p. 472). In probability methods, panelists may be overly confident about their superficial judgments. For example, panelists may respond “yes” with assuredness to an item that requires, for example, knowledge and skills that are included in PLDs but not in the borderline test-taker description. When making matching judgments, panelists may feel sure of matches between item knowledge and skill demands and the knowledge and skill expectations in PLDs, despite inadequate understanding or consideration of the demands and expectations of the two entities to be matched.

The *WYSIATI effect* (Kahneman, 2011, pp. 85–88) occurs when we base judgments on what we can represent in working memory without evaluation of the quantity or quality of the evidence. Panelists may make probability judgments based on attention only to selected knowledge and skills in the borderline test-taker description. Or they may fail to extrapolate beyond the borderline test-taker description to cover the specific knowledge and skill response demands. In matching judgments, panelists may evaluate collections of evidence, item knowledge and skill response demands, and score profiles based on limited consideration of those demands that are easy to call to mind, such as surface features of the items (e.g., vocabulary requirements) or the knowledge and skill requirements in the PLDs.

Confirmation bias happens when we locate and use information (e.g., recollections) to preserve a belief and disregard disconfirming information (see Ross & Anderson, 1982, p. 149). In probability methods, panelists may make judgments about borderline test-taker responses to items based on their individual beliefs about borderline test takers—that is, not the shared description—and their judgment of the cognitive challenges in responding to an item (i.e., the affect heuristic); they may support these estimates in subsequent rounds of judgments by selecting only supporting views offered by other panelists. In matching methods, panelists may weight features of collections of evidence, item knowledge and skill response demands, and score profiles and skill expectations in descriptions of levels of performance to justify the matches they prefer.

When the *affect heuristic* is operating, we allow likes and dislikes to determine our judgments rather than evidence-based analytic judgments (see Kahneman, 2011, p. 103). In probability methods, panelists may alter their judgments for some items to achieve a desired cut score (or alter an undesired cut score in subsequent rounds), rather than implement the intended cognitive-judgmental task. Standard-setting facilitators have observed that panelists tend to prefer and even work toward cut scores

that are two thirds of the total possible test score (see Davis-Becker et al., 2011). In matching methods, panelists may match collections of evidence, item knowledge and skill response demands, and score profiles with PLDs based on their likes and dislikes for particular pieces of evidence, items, content area subdomains they consider too difficult or easy, specific score profiles, or desire for a specific location of a cut score.

We employ *anchoring and adjustment bias* when we estimate an initial value and then make inadequate adjustments toward accuracy (Tversky & Kahneman, 1974, pp. 427–430). In both probability and matching methods, panelists may constrain probability judgments for some items in Round 1 based on their judgments to previous items, perhaps because they appear similar in knowledge and skill response demands. In addition, their Round 2 judgments may remain anchored to Round 1 judgments, despite contradicting insights offered by other panelists during feedback and discussion.

The *halo effect* occurs when judgments of one feature of an entity influence judgments about other features. People cannot easily evaluate individual features of people or objects fully independently of their evaluations of other features; as a result, they produce correlated ratings (Plous, 1993, p. 45). Horn effects produce the opposite of halo effects. Primacy, recency, and other ordering effects influence the strength of recall in working memory of multiple entities. The contrast effect can lead to inordinately higher judgments of a slightly better entity after a string of mediocre entities and vice versa (see Daly & Dickson-Markman, 1982; Plous, 1993, pp. 42–44). Panelists may make probability judgments for some items based on their previous judgments of items that appear to be similar in terms of knowledge and skill response demands. The influence of the halo effect appears similar to the effects of the representativeness and recognition heuristics, even though they have been researched separately in the judgment and decision-making literature and thus may appear to be different cognitive processes. Similarly, panelists may match pieces of evidence, item knowledge and skill response demands, and score profiles to PLDs based on their previous judgments that appear to be similar in terms of knowledge and skill response demands.

Therefore, panelists may be influenced by the order in which items are presented. Some panelists have reported that they work their way through all items and then review and change judgments for some earlier items, whereas others indicate that their comfort and understanding of the task increases as they work further through the test (i.e., they have more confidence in later judgments than in earlier ones).

WORKSHOP DESIGNS, PROCESSES, AND PRACTICES: MITIGATING UNDESIRABLE COGNITIVE AND SOCIAL INFLUENCES STANDARD-SETTING WORKSHOP DESIGN AND EXECUTION

The cognitive and social model of standard setting (Figure 12.2) proposes a comprehensive view of the standard-setting process and sources of influence on panelist judgment and decision-making. As shown in this model, the outcome of

standard-setting activities is influenced by a number of factors, many of which are determined during workshop design and execution. Earlier, we described the workshop process submodel (Figure 12.1), which explicates steps that are designed to immerse panelists in the standard-setting process of a specific standard-setting method; help them develop shared understandings of the required concepts, criteria, and procedures; and lead them through two or more rounds of making standard-setting judgments. In this section, we review the key decisions related to the design and execution of standard-setting workshops and propose strategies that standard-setting designers and facilitators can build into training with the goal of debiasing panelists. We base these strategies on what we have learned through our review of the cognitive-judgmental tasks within common standard-setting methods and the potential for cognitive biases and heuristics to impact how panelists apply the tasks.

Workshop Design

The design details of any standard-setting study are specified through a series of decisions made during the design phase. However, designers and facilitators may make adjustments during execution of the workshop. These decisions reflect all aspects of the cognitive and social model of standard setting because they include selection of the method (and the corresponding cognitive-judgmental task), selection of the panel, and design of the process.

A standard-setting *designer* makes key design decisions as to how standard(s) will be set. A standard-setting *facilitator* executes the standard-setting workshop and works directly with panelists. In some cases, the same person may serve these two roles, but for the purposes of this chapter we describe the responsibility of each separately. In this section, we outline key design decisions including selecting the standard-setting method, content or work samples to be used in the standard setting, empirical information to inform the process, the expert panel, and the meeting mode.

Selection of a Method

Method selection begins with identifying options that are feasible, based on the characteristics of the testing program (e.g., item formats, psychometric analysis), availability of required elements (e.g., test-taker work samples, test-taker performance data), and the appropriateness of the method's cognitive-judgmental task. Our analysis of cognitive processing requirements and cognitive-judgmental tasks of different standard-setting methods (see Table 12.2) suggests potential challenges for justifying selection of some methods. For example, the judgment and decision-making literature indicates that people generally do not make accurate probability judgments (e.g., Murphy, 2004) and percentage estimates (Gigerenzer & Hoffrage, 1995, as cited in Larrick, 2007, p. 325). In addition, some methods, such as Angoff, entail relatively complex cognitive loads. The famously provocative "fundamentally flawed" quote (Shepard et al., 1993, p. 77) refers

to the cognitive complexity of Angoff probability judgments (see also Pellegrino et al., 1999). In addition, as the field evolves toward writing technical reports that are oriented toward claims, evidence, and validity arguments (e.g., Ferrara et al., 2023; see also Kane, 2013), supporting the selection of some standard-setting methods may become even more problematic. Therefore, choosing a method that minimizes cognitive complexity, cognitive load, and other cognitive-judgmental task requirements may have to supersede past practice.

Selection of Test Content or Work Samples

The specifics of the selected method dictate what type of content needs to be a part of the standard-setting process. We use the term *content* to refer broadly to items, test form(s), work samples, or scoring rubrics. The test content reviewed in a standard-setting process should be selected so that the findings can be generalized to other forms of the test and other test-taker groups. Standard-setting studies typically focus on one form of a test and the resulting cut score(s) are equated to other forms. In some cases, test forms are quite long and the task of reviewing and judging an entire test form can become tedious.

Various studies (e.g., Coraggio, 2008; Ferdous, 2005; Kannan, Sgammato, & Tannenbaum, 2015; Kannan, Sgammato, Tannenbaum, & Katz, 2015) have evaluated the reasonableness of using a subset of items to set cut scores using an Angoff method. Results from early simulation studies (Coraggio, 2008; Ferdous, 2005; Ferdous & Plake, 2005, 2007) support the use of subsets of items. Resampling of operational standard-setting data was used to determine that approximately 45 items are sufficient to obtain generalizable cut scores (regardless of the length of the full test; Kannan, Sgammato, Tannenbaum, & Katz, 2015). This finding was then confirmed in an operational study where the results from two panels (one evaluating a 45-item subset and the other a full-length test) were comparable (i.e., within one standard error; Kannan, Sgammato, & Tannenbaum, 2015).

For item-mapping methods like Bookmark and ID Matching, a single test form may not cover the entire score scale with sufficient differences in item difficulty to facilitate judgments (Lewis et al., 1996). Therefore, creating an OIB may require sampling additional items from the bank to faithfully represent increments on the score scale, fill in large gaps between item scale locations, and reduce clustering of item locations, especially in ranges of the scale where cut scores could be located. This makes for a longer judgmental process (i.e., more items to review) but ensures panelists can make a cut score recommendation at the precise location on the score scale that corresponds with their expectations.

Other methods require panelists to set expectations for performance by matching work samples to PLDs (e.g., Body of Work, Judgmental Policy Capturing, and the Analytical Judgment methods). Although each of these methods is different, they all require study designers to preselect work samples for review. The work samples should

have been consistently scored and have good person fit (Kingston & Tiemann, 2012). Reviewing scoring rules/guidelines as well as work samples enable panelists to (a) understand the knowledge or skills required to achieve a given score for methods such as the extended Angoff (designed specifically for polytomous items), and (b) be able to match work samples to PLDs based on the scores on a rubric in methods such as Analytical Judgment.

Selection of Empirical Information as Feedback to Panelists

Standard-setting designers must also determine what empirical information will support panelists in their judgmental process. We identify three types of feedback that may be presented to inform panelists judgments: results from their judgments, item difficulty, and impact data. A later section on workshop execution details how feedback should be presented in a standard-setting workshop.

JUDGMENT RESULTS AS FEEDBACK After each round of judgments, facilitators have the opportunity to help panelists see how their judgments translate into a recommended cut score and how their judgments compare to those of others from the panel. Reckase (2001) referred to this as “process feedback” (p. 165) because it may encourage panelists to reconsider their judgments and increase the likelihood of convergence of their recommendations. This may include providing panelists (a) their individual recommended cut scores, (b) the range and average of panelists’ recommended cut scores, and/or (c) a full distribution of all panelists’ recommendations.

ITEM DIFFICULTY AS FEEDBACK In some standard-setting methods, panelists are asked to either explicitly (e.g., Ebel) or implicitly (e.g., Angoff, Bookmark) evaluate the difficulty of items or tasks by estimating probabilities or percentages as part of the cognitive-judgmental task. The empirical item difficulty may be presented as part of the materials (e.g., within an OIB in the Bookmark method) or provided later in the process (e.g., Angoff) for panelists to evaluate their judgments and guide refinement of judgments in later rounds. This feedback may be presented as item p values, IRT scale difficulty locations, or average score (for multipoint items). Reckase (2001) also classified this as process feedback. Reckase and Chen (2012) offered other, complex representations of item difficulty information (e.g., by test-taker score, by content area). It is important that the facilitator ensures panelists understand how to interpret difficulty information and incorporate this information into the cognitive-judgmental task as they make their decisions.

IMPACT DATA AS FEEDBACK Most standard-setting methods focus on setting expectations that represent the expected knowledge and skills of borderline test takers. However, providing impact information (e.g., pass rates, percentages of test takers classified within each performance level) can serve as a reality check (Green et al., 2005; Reckase & Chen, 2012) and help to assess the reasonableness of these recommendations.

Reckase (2001) referred to this as “normative feedback” (p. 163). Impact data are presented either after each round of standard setting or just before the final round. If available, the estimated impact of recommended cut scores can be evaluated by comparing to the impact from external benchmarks (e.g., state, national, and international tests; see Ferrara et al., 2021).

When data from operational administration are not available, one option is to provide historical impact data. However, historical data may represent a different exam (e.g., different content framework or format; Reckase & Chen, 2012). Another option is to examine the impact data from a similar exam (e.g., same or similar population and content focus).

In some cases, prior to the standard-setting meeting, the policy body may meet to discuss practical expectations for the standards to be set. This discussion may go as far as to set expectations for a range of standards or impact that would be tolerable, based on the information available (e.g., performance expectations, historical data). If available, these tolerance ranges may be provided to standard-setting panelists, in the form of target cut scores or benchmarks (see Ferrara et al., 2021) to help inform their judgments.

Empirically based approaches to standard setting (e.g., Ferrara et al., 2021; McClarty et al., 2013; O’Malley et al., 2012) are unique in that impact data are central to framing the judgmental task for panelists in the form of target cut score ranges. We opened this chapter by recognizing that standard setting is policymaking. Some designers take the point of view that impact data are relevant to policy decisions and should not be in the hands of content experts who are standard-setting panelists (e.g., Ferrara et al., 2021). When using an empirically based approach, impact is central to the cognitive-judgmental task and therefore must be reliable and well understood.

INFLUENCE OF DATA ON PANELISTS’ IMPLEMENTATION OF THE COGNITIVE-JUDGMENTAL TASK The research in this area is very clear about the value of data. While it can help to increase convergence among panelists (Margolis & Clauser, 2014) and increase panelists’ confidence in their judgments, it also influences their judgments (Brandon, 2004; Clauser et al., 2013; Impara & Plake, 1997; Margolis & Clauser, 2014; Mee et al., 2013; Skorupski & Hambleton, 2005). Specifically, panelists substantially modify their judgments irrespective of whether the data are accurate (see Clauser, Mee, et al., 2009; see the WYSIATI effect in the previous section). Similarly, when the estimated impact is very different from what a panelist would expect, it will likely drive the direction of adjustments to their judgments in the subsequent rounds of the standard setting (e.g., see the affect heuristic; and anchoring and adjustment bias).

Overall, presenting any type of performance data to standard-setting panelists requires caution and thought. First, designers should consider the relevance of the data to the intended interpretation and use of the cut scores. Knowing that panelists trust data even if they are inaccurate, it is important that the data serve as a reliable estimate of item difficulty or impact on the target test-taker population. If the only available

data are based on a different group of test takers or from a small sample of the target population, standard-setting designers should question the value of providing this information. Second, the ability of panelists to comprehend and interpret the information should be considered. Some item-level data are more straightforward than others. For example, a p value is more directly interpretable than IRT difficulty estimates (i.e., b values), and especially for polytomously scored items. Finally, policy makers or study designers may not support providing this information (Hambleton & Pitoniak, 2006), since the introduction of such information may alter panelists' judgments to the point where they no longer focus on the target cognitive-judgmental task. Therefore, standard-setting designers should consider the possible sources of this information and how it may help in the standard-setting process while enabling panelists to remain focused on the cognitive-judgmental task.

Expert Panel Formation

Standard setting is a judgmental process. Therefore, selection of a representative sample of appropriately qualified experts to carry out these judgments is critical to the overall validity argument for standards recommended in a workshop (Raymond & Reid, 2001). More specifically, because standard setting should be managed as a convergence process, diverse points of view and independent judgments are the key to achieving the most appropriate group recommended cut scores. Galton demonstrated the importance of these two features by observing expert farmers guessing the weight of an ox at the West of England Fat Stock and Poultry Exhibition at Plymouth, England.⁶ He observed that the *vox populi* (i.e., the median; Galton, 1907, p. 450) was within 1% of the correct weight while the individual estimates were up to 4% distant from the correct weight. The median recommended cut score in a standard setting is the best or most appropriate indication of a group's recommendation because it reflects the best combined judgments of all of the expert panelists and adjusts for highly discrepant judgments. Surowiecki (2005) resurrected Galton's observation and popularized the term *the wisdom of crowds*, which helps illustrate the importance of diversity and independence in standard-setting judgments.

PANEL COMPOSITION Based on the intended score interpretations (i.e., claims to be made) and uses (i.e., decisions to be made), standard-setting designers can identify qualifications for individuals to serve as subject matter experts. The definition of a subject matter expert can differ based on the context and purpose of the examination (e.g., entry level vs. advanced certification; Cizek & Bunch, 2007). At a minimum, individual panelists should have content area expertise (Hambleton & Pitoniak, 2006; Raymond & Reid, 2001) and be familiar with the target test-taker population to envision how a test taker would interact with test content. In educational testing, a standard-setting panel for an educational assessment may be composed of content experts such as teachers (who focus on the target content area and test-taker population) and nonteacher

educators (e.g., content specialists, administrators). In credentialing, a standard-setting panel is typically composed of veteran practicing professionals, recently credentialed professionals, and educators/trainers. Some panelists may consider themselves specialists (e.g., they hold the general credential but work within only one sector of the defined field) but are expected to be able to make judgments about all knowledge and skill expectations for new practitioners. In both education and credentialing, the panel should represent most (if not all) major subsets of the larger test-taker population.

For some contexts, non–subject matter experts who have a familiarity or a vested interest (Kane, 2001) with the target test-taker population may be included in the panel. They may be members of the public or those who have a vested interest in the decisions to be made based on the outcomes of the tests (Kane, 2001). Examples include parents of children taking the educational assessments, professionals who work collaboratively with those who hold the credential being sought, or members of the public. However, the inclusion of nonexperts has been questioned (Plake, 2008) because they are likely to lack the specific content knowledge and familiarity with the test-taker population to develop the shared understanding required to complete the judgmental task as intended and recommend defensible cut scores.

Panel diversity varies based on the type of program and may include personal characteristics (gender, race, ethnicity, or age) and professional factors (years of experience, credentials, type of employer, job role, and geographic location). In education, designers also should consider the diversity of the test-taker population who will take the assessment (e.g., general education, advanced classes, English language learners, students with disabilities) and the types of students represented by each subject matter expert. In credentialing, the panel should also represent the diversity of professional practice, including area of focus, professional responsibilities, and populations served by potential panelists.

Panel diversity may also include the perspectives (and possibly goals) that panelists bring to the standard-setting activities. For some panelists who are educators, the standard setting may have a direct impact on their own students or their school or district (i.e., in school and district public accountability reporting). In credentialing, some practitioners may feel inclined to set higher standards than necessary to limit competition within the field and hold candidates to a higher standard, whereas educators may feel inclined to set lower standards to increase the likelihood that their students will earn the credential (Buckendahl & Davis-Becker, 2012). All of these can result in misclassifications and violate the expectations in the *Standards* (AERA et al., 2014; see Standard 11.16).

Before selecting a panel, standard-setting designers are encouraged to set targets for the representativeness of the panel (e.g., Cizek, 2022; Hambleton et al., 2012) that reflect the characteristics of the larger test-taker and candidate populations. This requires casting a wide net to recruit experts that represent various strata of test takers and candidates. Based on the influence of group interactions within the cognitive and social

model of standard setting, we recommend looking beyond these traditional markers of surface-level diversity (e.g., adequate representation of race and gender) to consider deep diversity (Harrison et al., 1998), which includes heterogeneous attitudes, beliefs, and values that would be appropriate to represent in the panel. While it is likely not feasible to obtain information about these deep diversity characteristics from panelists ahead of recruitment, the degree to which deep diversity is achieved could be evaluated using questionnaires about the attitudes, beliefs, and values of panelists and evaluating those against the values believed to matter most for achieving performance standard recommendations based on diverse points of view.

NUMBER OF PANELISTS But how many panelists are enough? There is no set number for all standard-setting contexts. Standard-setting designers should include enough panelists to (a) achieve a reasonable representation of relevant expertise in the field, (b) yield small enough cut score standard errors to support the interpretation of cut scores, (c) enable full group and subgroup collaboration, and (d) include diverse opinions so that one perspective does not dominate panel-level discussions. Panels of 15 to 20 experts typically are recommended for state-level educational assessments (Hambleton et al., 2012) to ensure panel diversity and representativeness. In practice, 10 to 15 panelists have been found to yield acceptable standard errors (Raymond & Reid, 2001) and can be more manageable during panel-level discussions. Standard-setting designers should balance the need for enough panelists to ensure diversity and manageable size to be able to elicit all points of view. In credentialing, panel sizes of 10 to 15 panelists are more typical for large-scale programs (e.g., national licensure programs such as nursing, architecture, engineering). For smaller credentialing programs (e.g., highly specialized certification such as specialized technology fields, medical subspecialties), fewer experts may be available within the broader population and therefore smaller panels of 7 to 8 often are used in practice. The recent trend in standard setting for educational achievement tests is toward smaller panels of 8 to 10, primarily because recruiting teachers out of classrooms for multiple days can be challenging. Multiple, parallel, independently operating panels can enhance the generalizability of recommended cut scores and will often produce consistent results (e.g., Tannenbaum & Kannan, 2015). Some workshops include the use of a single panel with multiple subgroup tables for discussion. This enables richer discussion, more participation by each panelist, and comparisons of cut scores across subgroups to estimate cut score standard errors and evaluate consistency without having to run concurrent parallel panels. Generalizability analyses can be applied to evaluate a variety of assumptions about numbers of panels, panelists, and items to be used in a standard-setting study. For example, Hurtz and Hertz (1999) used a generalizability study to determine the optimal number of panelists necessary for standard-setting studies. In evaluating eight occupational licensing examinations, they found that 10 to 15 panelists would be optimal, based on the reduction in expected standard errors.

Remote Versus In-Person Workshops

Traditionally, most standard-setting workshops involved face-to-face meetings. However, some programs successfully conducted virtual standard-setting processes even before remote meetings were commonplace (see Katz et al., 2009; Schnipke & Becker, 2007; and Shaw, 2007, for some early implementations of virtual standard-setting meetings.) The Smarter Balanced Assessment Consortium (2015) was the first high-stakes educational program to implement a fully online process to gather crowd-sourced Bookmark judgments from over 2,000 educators from a number of states.

Advantages of virtual standard setting largely are logistical (Schnipke & Becker, 2007) because they are cost-efficient (no requirements for travel, meeting space, or meals) and only require participation (not travel) time (e.g., Katz et al., 2009). In turn, programs may be able to recruit larger, more representative panels for virtual meetings. Comparative research on in-person versus virtual standard-setting workshops generally indicates similar recommended cut scores (Harvey & Way, 1999; Katz & Tannenbaum, 2014; Kollias, 2018) and reports of positive experiences from participating panelists (Katz et al., 2009; Kollias, 2018).

However, some concerns remain about conducting a high-stakes process in this manner. First, virtual meetings leave more opportunities for panelists to be distracted and less engaged than in-person meetings. Although facilitators can monitor for engagement (e.g., track input such as comments and questions) and use encouragement strategies, less group collaboration may impact the panel's ability to develop a shared understanding of the task and expectations on which to base their judgments. Second, working face to face provides more opportunities for a facilitator to evaluate panelists' understanding of the task (e.g., evaluating body language) and offer additional explanation and training (e.g., one-on-one conversations). Third, distributing secure test content outside a secure meeting room increases the risk of compromise. Such risks are being increasingly managed with improvements in virtual collaboration tools and remote meeting technology. As virtual meetings become more commonplace, these concerns may be reduced as a result of familiarity with this format and improved virtual meeting technology that allows increased flexibility for addressing the needs of meeting participants.

Judgmental Rounds

Most standard-setting workshops are designed as an iterative process where panelists make judgments, receive feedback on their judgments and possibly additional empirical information, and then revise their judgments accordingly. Most standard-setting designers would advocate that multiple rounds of standard-setting judgments enable panelists to review and revise their judgments after going through the process entirely (i.e., Round 1), learning how their judgments are transformed into a cut score, considering how their judgments compare to those of the panel, and reviewing empirical information. However, there is debate as to whether more than two rounds are necessary. Some advocate that additional rounds allow facilitators to layer in feedback (e.g., results

and process feedback provided first, followed by any empirical or normative feedback), whereas others suggest that a third round simply reduces variability in the results due to continued discussion while adding significant time and expense (Cizek, 2022).

Validation Framework to Guide Workshop Design

In concluding this section on workshop design, we turn our attention to validity evidence that is collected throughout the workshop and used to form validity arguments for the defensibility, generalizability, and intended interpretations and uses of cut scores. Building off previous guidance offered by Hambleton and Pitoniak (2006) and Kane (1994, 2001), we propose a framework of expectations and supporting evidence aligned with the claims presented earlier in the chapter that:

1. The cognitive-judgmental task (inherent in the standard-setting method) is a critical aspect of standard setting and should be central to the design, preparation, execution, and evaluation of the workshop.
2. In addition to the cognitive-judgmental task, there are other influences on panelist judgment and decision-making in standard setting. Some are design features and are desirable; some are not.
3. Standard-setting results should reflect a well-executed process that was led by experts who helped panelists make their judgments avoiding undesirable influences.

In Table 12.3, we describe the validity expectations and suggested documentation for each type of evidence: design, preparation, decision-making, and outcomes evidence. Because validation begins with design and design of the workshops influences the recommended cut scores, the expectations laid out in this framework should be considered in making the workshop design decisions.

Workshop Process

In this section, we focus on workshop processes (i.e., workshop process submodel in Figure 12.2). In addition, we review how final cut score decisions are made and how the entire process is evaluated and documented. Facilitator processes in the group interactions submodel are implicit in each of these steps.

Getting Oriented: Building Familiarity With the Test and Performance Expectations

Several activities within most standard-setting workshops are devoted to orienting the panelists to the standard-setting context and forming some “key conceptualizations” (Cizek & Bunch, 2007, p. 48). Orientation typically includes activities such as familiarizing panelists with the test and discussing expectations for performance on the test (e.g., PLDs). Orientation activities typically are conducted before training panelists in the standard-setting method’s cognitive-judgmental task.

Table 12.3 Framework to Guide Workshop Design and Gather Validation Evidence to Support Claims about the Interpretation of Cut Scores

Validity Expectation	Suggested Documentation
Design Evidence	
The study design aligns with the purpose and characteristics of the test and cut score(s) to be set.	
1. The standard-setting method and its cognitive-judgmental task are suited to the format of the test, expertise of the panelists, and the intended interpretation and use of test scores.	- Rationale for the selected standard-setting method and how the cognitive-judgmental task is appropriate for the panelists and intended interpretation and use of test scores.
2. Empirical information presented during standard setting (e.g., external benchmarks and target cut scores, item difficulty, performance distributions) is consistent with the standard-setting workshop design and method.	- Description of the empirical information including the sample of test takers represented and when it was collected. - Details as to when and how the empirical information was presented during standard setting, including instructions provided to panelists to guide interpretation and inform judgments.
3. The panel collectively represents the range of expertise needed to inform the cut score recommendations.	- Panelist qualifications such as employer, position, years of experience, geographic location, and populations served.
4. Standard-setting activities (i.e., training, judgment activities, feedback), ways of integrating expert judgment, empirical information, and discussion are specified in the workshop design.	- Details about the standard-setting plan that includes the inputs to, and collection of, judgments used to determine the recommended cut score(s).
Preparation Evidence	
Panelists are adequately prepared to make the necessary judgments.	
5. Panelists developed a shared understanding of the item response demands and expectations for each performance level.	- Details about the review of PLDs and/or the development of threshold PLDs. - Panelist evaluation responses indicating they had a sufficient understanding of the expectations for each performance level and/or the threshold performance expectations.
6. The training adequately prepared panelists for the standard-setting task, including following the cognitive-judgmental task and addressing potential heuristics, biases, and other factors that can distort decision-making.	- Details about the training activities including topics covered and specific instructions provided to panelists including what to think about and what shortcuts to avoid. - Panelist evaluation responses indicating they felt adequately prepared to follow the cognitive-judgmental task.

(continued)

Table 12.3 (*continued*)

Validity Expectation	Suggested Documentation
Decision-making Evidence	
Panelists make standard-setting judgments following the standard-setting design and plan.	
7. Panelists applied the method's cognitive-judgmental task as intended, including consideration of additional feedback information that was provided.	- Panelists' evaluation responses indicating they followed the cognitive-judgmental task during each round of judgments.
8. The process provided opportunities for panelists to receive and discuss meaningful feedback on their judgments and revise their judgments, if warranted.	- Panelists' evaluation responses indicating that the feedback and discussions were valuable to inform their judgments.
9. Workshop facilitators managed processes without unduly influencing judgment and decision-making about cut score recommendations.	- Panelists' evaluation responses indicating that they have confidence in their recommended cut scores.
Outcomes Evidence	
The results show consistency in panelist expectations and the results are reasonable. ⁷	
10. The results indicate the panel's shared understanding of the knowledge and skill expectations for each performance level.	- Empirical information on the similarity in recommended cut scores among panelists (e.g., range, standard error).
11. The results indicate that panelists were consistent in their application of the cognitive-judgmental task.	- Empirical information that shows consistency between panelists' judgments and other measures of the item response demands (e.g., targeted performance level, empirical item difficulty).
12. The results are reasonable, based on external information or review by stakeholders.	- External evaluation of results that indicate reasonableness which may be empirical (e.g., comparison to historical results, results from another measure, results from another standard setting method) or subjective (e.g., review by policy body or subject matter experts).

FAMILIARIZING PANELISTS WITH THE TEST Various documents may be used to familiarize panelists with the test, the types of assessment activities (i.e., items), and the relative emphases placed on content strands within the test. These documents range from high-level descriptions of test content to more detailed test specifications or sample test forms for panelists to review or take (Cizek & Bunch, 2007). Whether panelists take the entire test or a selection of items, they then can discuss how the tested content matches the test specifications (Morgan & Michaelides, 2005). This enables panelists to develop a shared understanding of what the test measures. During this discussion, the facilitator's responsibility is to focus discussion on the *actual* tested content, clar-

ify any misperceptions about difficulty or fairness of the test, and remind the panelists that their task is to recommend cut scores on the test as it is presented to them, not to evaluate the content of the test.

Education standard-setting panelists may have some familiarity with the test but it is important to review this information to ensure everyone has the correct understandings. Conversely, participating in a standard setting may be a credentialing panelist's first experience in working on a test. Therefore, training for a credentialing standard setting usually involves a significant amount of time to orient panelists to the test's purpose, design, format, and general development process, as well as the purpose of standard setting in this process.

UNDERSTANDING AND DEFINING EXPECTATIONS Cut scores are set on tests to enable interpretation of performance in relation to expectations for leveled categories. It is important that standard-setting panelists understand the knowledge and skills expected at each performance level and the key differentiators between adjacent levels.

In credentialing testing, there are usually two performance categories, representing qualified and nonqualified candidates. Therefore, the focus of expectations is on the point of differentiation between these two levels, or the candidate who is on the borderline (i.e., minimally qualified, minimally competent). Panelists can develop a shared understanding of this minimum level of expectation by developing or reviewing the description of the borderline candidate. Expectations should be framed around anticipated job responsibilities for an entry-level professional, rather than what is emphasized in an educational program or experience opportunity (e.g., an internship).

In educational testing, PLDs and their labels represent the state's educational aspirations for its school-age children. The cognitive-judgmental task for most of the probability and some matching methods focuses on borderline test takers and may require the development of borderline PLDs. Range PLDs (Egan et al., 2012) define the expectations, within a performance level, of what individuals in a performance level should know or be able to do in a test content area (Cizek & Bunch, 2007). Cut scores reflect the minimum of knowledge and skills required to enter that performance level or to be licensed or accredited in a field (e.g., bare minimum knowledge and skills to be licensed as a teacher—not an expert teacher, but someone who is qualified enough to receive the keys to the classroom). Therefore, developing well-articulated PLDs is one of the most critical steps in arriving at reasonable cut scores. This importance has been particularly recognized in recent years (e. g., Bejar et al., 2007; Egan et al., 2012; Hambleton et al., 2012; Huff & Plake, 2010; Huff et al., this volume; Luecht, 2020; Perie, 2008).

It is now more common for these expectations (e.g., borderline test taker, PLDs) to be developed by separate panels prior to the standard-setting workshop because they are used in other test development activities (Bejar et al., 2006, 2007). Developing

descriptors ahead of a workshop can reduce panelist meeting time and expense because reviewing and internalizing descriptors can take less time than generating them. However, we have observed that going through the development process helps panelists internalize these expectations as they debate the expectations for a given level.

Learn: Training, Modeling, and Practice to Achieve Competence in Implementing the Cognitive-Judgmental Task

Standard-setting panelists are recruited because they possess content expertise and may represent a constituent stakeholder group. However, they cannot be expected to have experience with the chosen standard-setting method. Therefore, building in sufficient time to train panelists in the standard-setting method is critical so that they may appropriately carry out the cognitive-judgmental task to recommend a cut score (Cizek & Bunch, 2007; Hambleton et al., 2012). We define training as (a) presentation of concepts and the cognitive-judgmental task, (b) modeling of judgments and decisions, and (c) guided practice and focused feedback in implementing the intended cognitive-judgmental task.

Panelists should receive step-by-step instructions on the materials they will use and the factors they must consider as they complete and record their judgments (Hambleton & Pitoniak, 2006). First, panelists should be trained to focus on the correct target for their judgments (e.g., knowledge and skills of borderline test takers vs. the entire performance level) to properly implement the intended cognitive-judgmental task. Second, panelists should be explicitly trained in implementing the intended cognitive-judgmental task for the chosen standard-setting method. Third, they should learn about the heuristics, biases, and other negative influences that may arise for the selected method. Finally, after the cognitive-judgmental task is clearly described and modeled, panelists should have the opportunity to practice (e.g., Plake, 2008) and receive guided feedback. This practice activity allows panelists the opportunity to solidify their understanding of the cognitive-judgmental task and the descriptions of the borderline test takers developed/reviewed in the training.

The materials used for practice (e.g., test items, test-taker responses, scoring keys and rubrics, descriptions of borderline test takers) should be similar to those panelists would use in the actual task. To avoid biasing subsequent independent judgments, there should be no overlap of items or test-taker responses used for practice with those that will be used in the actual rounds of judgment. Facilitators can uncover any misunderstandings about the intended cognitive-judgmental task or underlying biases during the practice discussions. We know from several studies (e.g., Davis-Becker et al. 2011; Hein & Skaggs, 2009) that some panelists may use the substitution heuristic (Kahneman, 2011) in undetectable ways to simplify their judgmental task. Facilitators can use these practice discussions to reiterate the purpose of the standard-setting study and refocus panelists if they see panelists straying from the intended cognitive-judgmental task.

Typically, panelists are asked to declare their readiness to proceed at the end of training using a self-report survey (i.e., evaluation). Panelist readiness also can take the form

of a quiz to evaluate their understanding of the intended cognitive-judgmental task. Self-report evaluations typically include statements about participants' understanding of the assessment and its purpose, the standard-setting process, the clarity of the borderline test-taker descriptions, the training and modeling provided in the cognitive-judgmental task, sufficiency of practice opportunity, and their overall comfort and readiness to proceed to implement the intended cognitive-judgmental task. Several authors provide sample evaluation forms that can be customized for any standard-setting workshop (e.g., Cizek, 2012b; Cizek & Bunch, 2007; Hambleton et al., 2012; Zieky et al., 2008). Effective training is essential to support the validity of the entire standard-setting process (Morgan & Michaelides, 2005) and the resulting cut scores. Yet, to our knowledge, no research has directly addressed the effectiveness of training on standard-setting panelists' ability to implement the cognitive-judgmental task while minimizing the influence of biases and unintended influences.

SOME PROPOSED TRAINING STRATEGIES FOR DEBIASING In an earlier section, we identified cognitive processes, judgmental heuristics, and cognitive biases that may influence the degree to which panelists follow the cognitive-judgmental task for some methods. In Table 12.4, we propose training strategies to reduce the deleterious effects of these biases and influences. Some proposed strategies are generalized from empirically supported training interventions. Others involve modeling cognitive behaviors, providing examples of thinking errors, and guiding panelist practice in overcoming judgmental errors. Our proposals are intended to help facilitators guide panelist thinking and decision-making—that is, to nudge them to make judgments and decisions with minimal interference from ineffective heuristics, biases, and other sources of errors. At the end of this chapter, we recommend research on the efficacy of these training strategies specifically in standard-setting situations and for specific standard-setting methods.

Feedback and Discussion

Typically, standard-setting processes include at least two rounds of judgments, with feedback and discussion between rounds. Impact data (i.e., normative feedback about test-taker performance) can influence panelists' cut score recommendations either directly or indirectly in methods like benchmarked standard setting (Ferrara et al., 2021). When there are more than two rounds of standard setting, Reckase (2001) advised the following: introduce criterion-referenced (e.g., item difficulty) feedback after the first round, followed by normative feedback (e.g., impact data) after the second round, with panel results provided at the end of each round. This enables panelists to refine their understanding based on criterion-referenced aspects (such as item difficulty) first, without the compelling influence of normative feedback, which is irrelevant to the fundamental cognitive-judgmental task.

One final form of feedback occurs naturally as panelists share viewpoints with each other between rounds. During discussions, panelists may explain their think-

Table 12.4 Proposed Proactive Training Strategies to Reduce Deleterious Effects on Panelist Judgment and Decision-Making of Judgmental Heuristics and Cognitive Biases (or “Debiasing”)

Overarching Cognitive Processes
Framing effects: State the cognitive-judgmental task exactly as panelists are expected to think and decide, ensure that panelists understand how they should and should not think, provide ample practice opportunities, and reinforce panelist understandings by asking them to restate and explain the statement.
Continuous monitoring and basic assessment process: Encourage panelists to self-monitor their thinking and decision-making as they implement the cognitive-judgmental task. Encourage panelists to raise questions and ask for guidance if they become uncomfortable with a part of the standard-setting process or a decision.
Overconfidence bias: Raising panelist awareness of the accuracy-confidence disconnect and reminding them to check and question their thinking and decisions may help to mitigate this bias. Personalized performance feedback on inaccurate or unfounded judgments and the “consider the opposite” strategy (Lord et al., 1984) have been shown to reduce overconfidence.
Judgmental Heuristics
Substitution: Ask panelists to restate the intended cognitive-judgmental task statement regularly. Discuss the importance of implementing the intended cognitive-judgmental task and avoiding substitution and other ways in which other panelists may work around the stated cognitive-judgmental task.
Intensity matching and values translation: There may be no effective way to train panelists to improve inconsistent and inaccurate translations of intensities from one scale (e.g., item response demands categories) to another (e.g., probabilities of a correct response). It does seem likely that a facilitator could ask panelists to rethink a judgment that is overheard as part of the monitoring process.
Availability: Reminding panelists that the ease of recalling category examples to mind does not always comport with the range of possible examples. Nudging them to think about other possibilities may help somewhat. In addition, encouraging panelists to listen to other panelists’ examples and taking the outsider view (i.e., considering broader implications rather than focusing only on the immediate situation) may be effective in calling to mind a wider range of possibilities.
Representativeness: Encouraging panelists to be disciplined and consider all response demands of items, requirements of performance-level descriptors (PLDs), characteristics of borderline test-taker descriptions, qualities of test-taker work samples, and achievement of known test takers may improve their matching judgments, but not probability judgments.
Cognitive Biases
WYSIATI effect: Raising awareness of this bias and encouraging panelists to consider all response demands of items, requirements of PLDs and borderline test-taker descriptions, qualities of test-taker work samples, and abilities of known test-takers may reduce this bias.
Confirmation bias: Training panelists to look for disconfirming as well as supporting evidence may be the most effective way to mitigate this bias.

Affect heuristic: Making panelists conscious of this heuristic, how it may influence their judgments (e.g., selecting a desired cut score rather than a supportable one), and nudging them to stay true to the intended cognitive-judgmental task may be the most effective intervention.

Halo and other effects: Mitigating the influence of adjacent and nearby items, work samples, and score profiles on a current judgment may not be feasible. In addition, it is likely that the contrast effect also may operate and perhaps cannot be mitigated. Raising awareness and urging discipline in making judgments and decisions may be the only recourse.

Anchoring and adjustment bias: Raising awareness of underadjustment in relation to Round 1 cut scores that serve as anchors for Round 2 judgments may free panelist thinking to deviate from their Round 1 decisions. However, significant deviation from Round 1 or from externally benchmarked anchors is inconsistent with the goal for subsequent standard-setting rounds (i.e., convergence of panelist judgments) and for benchmarked standard setting (i.e., retaining target cut scores or making only small adjustments to the anchors). Personalized performance feedback on inaccurate judgments may mitigate inadequate adjustments.

ing and consider others' explanations for their judgments. Social comparison (Festinger, 1954) may influence panelist thinking and judgments in subsequent rounds. Likewise, information sharing among panelists can lead panelists to adjust their judgments (e.g., persuasive arguments theory, Burnstein & Vinokur, 1975)⁸ and serves as feedback. The facilitator may provide guidance on persistent misunderstandings of concepts, the process, or indications of biases. Feedback obtained during discussions may also have substantial influence on the judgments made by panelists, as suggested by convergence of panelist judgments over rounds and reported in workshop evaluation surveys.

Postworkshop Activities and Documentation

FINAL CUT SCORE DECISIONS Standard-setting results (i.e., cut score recommendations) are determined based on the final standard-setting judgments. As an initial step in consolidating the results, standard-setting designers and/or facilitators can examine the results and explicitly describe the recommendations of the panel based on all the information collected during the workshop. In addition, standard-setting leaders can review workshop evaluation results and individual panelist recommendations for evidence that they followed the cognitive-judgmental task. If there is reasonable evidence to suggest that a panelist did not follow the intended cognitive-judgmental task (e.g., based on observations from the facilitator), standard-setting leaders may decide to exclude those judgments when calculating final cut score recommendations (i.e., "adjustments to participants"; see Cizek, 1996, p. 17). Taking such action is not intended to influence the overall recommendation in any particular direction or artificially decrease the standard error of panelist recommendations. Rather, this strategy enhances the defensibility of cut scores by using only those judgments that appear to follow the intended cognitive-judgmental task.

The final cut score recommendations are calculated based on the specific method. In many cases, the recommended cut score may be calculated as an average or median of

the panel's recommendations, which is typically presented within a range of cut scores that reflects the variability in recommendations across the panel. The size of the range is determined by the standard error (more variability = larger standard error). These results are then presented to the authoritative body/organization responsible for selecting the final cut score. Additional context may be provided to help policy makers interpret the results of the standard setting, such as estimated impact, historical impact, or comparison to other related measures. Buckendahl et al. (2005) described a process in which the estimated impact from the recommended cut scores (on a new statewide assessment) was presented in comparison to the previous state assessment and the state-level results from the National Assessment of Educational Progress (NAEP). Policy makers are advised to select the final cut score from within the recommended range given the defensibility of the standard-setting process and the validity evidence that supports the results.

In rare cases, an authoritative body will set the final cut score outside the recommended range. Although not often documented or shared, this typically occurs in one of two situations. The first is when a program is making a significant change to test design, structure, content, or format and the accompanying cut score represents a significant increase in expected knowledge and skills and a significant decrease in the pass rates. In such cases, a policy body may implement a phased approach, where a lower cut score is set as an interim step with the plan to increase the cut score (to the recommended level) over time. Policy bodies also may set a cut score different from the panel recommendation if they feel the panel did not provide them with a reasonable recommendation. This type of heavy-handed action may seem similar to a judge setting aside a jury's decision to enforce a different verdict in a courtroom setting. A policy body will consider such actions only if they have strong evidence based on observed conversations within the panel (e.g., collusion to advance a particular agenda beyond the intended focus of the standard setting) or empirical evidence indicating the cut score is not reasonable (e.g., expected impact of a 20% pass rate when similar measures have a 60% pass rate).

STANDARD-SETTING RESEARCH: CURRENT FINDINGS AND FUTURE DIRECTIONS

We conducted an extensive review of the empirical research literature on standard setting. We focus in this section on investigations into panelist experiences. In the broader review, we identified four areas of research on standard setting. These areas are:

- comparisons among standard-setting methods. Studies in this group compare two or more standard-setting methods, often focusing on the cut scores that are ultimately recommended.
- investigations of specific standard-setting methods. Studies in this group focus on a particular standard-setting method and common modifications to, or

assumptions about, a particular method. Studies proposing new methods with outcome data also are included in this group.

- explorations of panelist experiences. Studies in this group are concerned primarily with standard-setting panelists, what they think about as they make standard-setting judgments, and the effects of panelist characteristics on their judgments and decision-making processes.
- general issues in standard setting. Studies in this group address broader issues in standard setting that apply to different methods and contexts such as different modes (e.g., face to face or virtual) or how standard setting fits within the broader assessment field.

Given our emphasis on panelist cognitive processing and judgment and decision-making, we focus on the research in the third group above: explorations of panelist experiences. More specifically, we review findings from research that complement our proposed cognitive and social model of standard setting (Figure 12.2) and focus on what panelists think about when they participate in a standard-setting workshop. We organize this review into two sections: studies that examine (a) panelists' internal thinking about various components of the standard-setting process (reflecting the individual panelist cognitive processing submodel) and (b) external factors that are hypothesized to influence panelist thinking (reflecting the social and political context and the group interactions submodel). Our workshop process submodel is addressed implicitly because researchers frequently gather data from panelists about their experiences at key points during workshops through surveys or interviews. The review concludes with an explicit discussion of how these research findings relate to the model. Following the review, we suggest areas where further research is needed to better understand panelists' thinking and experiences during the standard-setting process.

Research on Standard-Setting Panelist Cognition

Research on Panelist Thinking in Standard-Setting Workshops

Research on panelist thinking during standard-setting workshops is concentrated among variations of the Angoff and Bookmark methods, which require panelists to employ probability judgments. Some research has examined how panelists conceptualize borderline test takers in these methods. Although this does not constitute a cognitive-judgmental task, panelist thinking about borderline test takers (and potential operation of heuristics and biases when doing so) has important implications for how panelists navigate the decisions they make.

PANELIST THINKING IN ANGOFF METHOD VARIATIONS Using focus groups, Ferdous and Plake (2005) investigated differences in panelist thought processes, based on panelists' propensity to provide systematically high, moderate, or low item performance estimates, using the modified Angoff method for fifth-grade reading and mathematics

assessments. They found that panelists who provided systematically high item estimates (resulting in higher proposed cut scores) thought about PLDs and their own students when making decisions. Panelists who provided systematically low estimates of borderline test-taker performance (resulting in lower proposed cut scores) also considered the policy impacts of cut scores in accordance with federal accountability requirements in *No Child Left Behind* (2001).

Skorupski and Hambleton (2005) examined panelist thought processes using questionnaires at key points in an Angoff yes/no standard-setting workshop for an English as a second language test. They focused specifically on the reading and speaking sections of the test. Panelists, all of whom were teachers, completed questionnaires immediately following the workshop welcome and orientation, the yes/no method training session, discussion of test content and PLDs, Round 1 judgments, discussion of feedback on Round 1 judgments, and following Round 2 judgments. Data from the welcome and orientation questionnaire revealed that some panelists did not understand the purpose of the standard-setting session; some believed they would be reviewing or scoring test items. Following their Round 1 judgments, panelists reported various strategies for making yes/no judgments, including considering how a reading item's difficulty could be affected by an associated reading passage and selecting "yes" when in doubt about borderline test takers, in both reading and speaking. Panelists generally reported that feedback between Rounds 1 and 2 was more helpful for the reading section than for the speaking section. Finally, following Round 2 judgments, panelists reported a greater focus on test content for the reading section than in Round 1 and reported not changing decisions between rounds for the speaking section.

Hansen et al. (2013) explored panelist experiences in an extended Angoff method workshop for three versions of a state's alternate science assessment for students with significant cognitive disabilities. Each assessment version differed in cognitive demands. The extended Angoff method was modified further by application of the mean estimation procedure, where panelists estimated average item scores for a hypothetical large group of borderline test takers (Loomis & Bourque, 2001). Following training, panelists in this workshop created PLDs for each version of the assessment before making their standard-setting judgments. They reported challenges with this process. For all three versions, panelists reported challenges in differentiating among high, medium, and low levels of performance. Several unique challenges emerged in crafting PLDs for the least cognitively complex version of the assessment, including "keeping 'low' definitions in positive terms" and being unsure "if enough science content was included" (Hansen et al., 2013, p. 309). For the more cognitively complex version of the assessment, panelists reported struggling to determine whether students at a low level of performance should be taking a less complex version (and vice versa for high levels of performance). Following their first round of judgments, panelists' greatest concern was that their judgments would be substantially different from the rest of the group. This concern dissipated after group discussion and the second round of judgments. Most panelists were confident in their judgments and applications of

the standard-setting procedure following the second round. Panelists also reported an array of factors important to their judgments, including the PLDs, group discussion, and perceptions of item difficulty.

Clauser et al. (2013) investigated panelists' use of different kinds of performance data in standard setting for medical licensure tests. In two modified Angoff workshops, the authors provided some panelists with limited performance data between rounds (i.e., the percentage of test takers selecting each response option) and other panelists with more extensive feedback (i.e., the response option information and test-taker performance on each item in deciles). They found that when panelists were provided more extensive data, their Round 2 judgments were more highly correlated with empirically estimated conditional item probabilities for borderline test takers than the judgments of panelists who were provided with limited performance data. However, they found no differences in how the different types of data impacted the cut scores that were ultimately recommended.

In an Angoff yes/no application for a licensure test, panelists recommended cut scores in three prespecified pass/fail conditions (Condition 1: where no target criterion was specified; Condition 2: with a 50% criterion; and Condition 3: with a 70% criterion). The Angoff yes/no judgmental task was stated as "Will a borderline test taker respond correctly to this item?" (In Conditions 2 and 3, the statement included "More than 50% of the time?" or "More than 70% of the time?") The average panelist cut score recommendations in Conditions 1–3 were, respectively, 60%, 50%, and 70% yeses, suggesting that when a criterion was specified, panelists may have altered their item-by-item judgments, on average, to approximate the specified criterion (R. Tannenbaum, personal communication, April 23, 2020). Framing effects, an overarching cognitive process, explain what may have operated here. The result also suggests that compliance was in operation (defined as "a change in a person's behavior in response to a direct request," (APA, n.d.).

PANELIST THINKING IN THE BOOKMARK METHOD Dawber et al. (2002) used survey and think-aloud procedures to probe panelist bookmark placement strategies and considerations for high school examinations. Across rounds, panelists commonly reported placing bookmarks through identification of a promising interval of items and selecting among the items in the interval or identifying a single promising item and "focus[ing] on that item and the preceding items" (Dawber et al., 2002, p. 36). They also found that panelists considered their own professional experiences with students, content standards, and PLDs in initial bookmark placements and expanded this list of considerations to include empirical item difficulties (when provided) and group discussion in their second round of bookmark placements.

Hein and Skaggs (2009) found that panelists in traditional and single-passage Bookmark⁹ standard settings struggled with selecting a single item and OIB page on which to place their bookmark. They also reported that several panelists deviated from the method's intended cognitive-judgmental task, particularly in the traditional Bookmark

condition. Like Dawber et al. (2002), they report that some panelists identified a range of items for bookmark placement and struggled to settle on a single item. Some panelists also reported calculating their bookmark placements mathematically, based on the point in the OIB where a particular percentage of content standards were reflected in the items. These qualitative reports complement the findings of other empirical work (e.g., Davis-Becker et al., 2011; Clauser et al., 2017) that suggest that panelists use alternative strategies to place bookmarks (i.e., they use the substitution heuristic). Hein and Skaggs (2009) found that panelists persisted in the use of alternative strategies even when redirected to the intended cognitive-judgmental task, defending their strategies as the pursuit of “adequate justification” for their bookmark placement (p. 219). These findings suggest that panelists may have engaged in algorithmic thinking, seeking rules to lead to what they considered the best bookmark placement.

In an investigation of a Bookmark standard setting for an education leadership exam, Cravens et al. (2013) reported that panelists found PLDs helpful in guiding their judgments, but relied on these descriptions less as they progressed through each round of bookmark placement. The authors suggest this may be due to increased panelist confidence and internalization of the PLDs over the course of the workshop. Panelists also reported appreciation of guiding questions provided during the process to keep the focus on the criterion-referenced nature of the task:

(a) What behaviors must a principal exhibit to achieve a rating of *at least highly effective* [i.e., a 4; italics in the original] on this item? and (b) What makes it more difficult to achieve a rating of *at least highly effective* on this item than on all previous items in this book? (Cravens et al., 2013, p. 134; italics in original)

Panelist thought processes also are addressed implicitly in several other studies. For example, Davis-Becker et al. (2011) found that, in a Bookmark standard setting, panelists placed their bookmarks in similar positions for a randomly ordered OIB and a difficulty ordered OIB, suggesting that panelists may be paying more attention to the percentage of items captured by their cut score rather than difficulty of content and relationship to PLDs. Clauser et al. (2017) obtained similar findings when they compared panelist bookmark placements in OIBs that were designed to target different levels of ability.

Similar findings also have been observed for the response probability (RP) criterion in the Bookmark method. Karantonis and Sireci (2006) suggested that the commonly used RP 67 is helpful in assisting panelists in conceptualizing mastery of test content. However, the selection of an RP criterion may have implications for proposed cut scores. Hambleton and Pitoniak (2006) suggested that, theoretically, different RP criterion should lead to equivalent performance standards, even if they result in different bookmark placements by panelists. However, some research disputes this theoretical principle. Wyse (2011) sought to investigate circumstances under which the use of different RP criteria would lead to the identification of identical cut scores and found that this is very unlikely to occur in practice. Baldwin et al. (2020) examined differences

in proposed cut scores when panelists were randomly assigned to consider response probabilities of .67 or .80 for identical tests. They found that panelists placed book-marks in similar locations in both conditions. This led to a higher cut score in the RP 80 condition, which resulted in a higher than projected failure rate.

PANELIST THINKING ABOUT BORDERLINE TEST TAKERS In two separate Angoff studies for fourth-grade reading and high school mathematics tests, where initial borderline test-taker definitions of varying specificity were provided, Giraud et al. (2005) investigated (a) correspondences between teachers' descriptions of *Barely Proficient* or *Barely Master* test takers and (b) factors they identified during training that might make assessment activities easy or difficult for these students. The fourth-grade reading panelists were provided with a relatively broad definition for a *Barely Proficient* student, while the *Barely Master* definition provided to the high school mathematics panelists was more specific. They found that panelists generated more varied descriptions of borderline test takers when the facilitator provided a less specific initial definition.

Panelists may conceptualize the behaviors of more than one hypothetical borderline test taker. Hein and Skaggs (2010) examined teachers' conceptualizations of classrooms of borderline students using the single-passage Bookmark and Angoff yes/no methods. Panelists reported difficulty with the prescribed cognitive-judgmental task for both methods, which involved imagining "an entire classroom made up of only barely proficient students" (Hein & Skaggs, 2010, p. 43). Panelists in the single-passage bookmark condition were instructed to place their bookmarks at the point where two thirds of the class would succeed on the item; panelists in the Angoff yes/no condition were instructed to select yes if they believed that at least two thirds of the class would succeed on an item. In both methods, panelists reported struggling not to think about their own classes, which are not composed solely of *Barely Proficient* students. They reported several alternative strategies, the most prominent for the Angoff yes/no method being that panelists thought about the borderline students within their own classrooms. Panelists using the single-passage bookmark method tended to focus on a specific borderline *Proficient* student from their own class and then imagine a classroom composed of other, similar students.

Research on Factors Hypothesized to Influence Panelist Thinking

Although not every component of our proposed group interactions submodel is addressed in the research literature (for example, readily accessible research on standard-setting facilitation is particularly sparse), topics such as panelist identity and group interactions have been explored to some degree. Some research also has examined the impacts of social and political contexts on panelist thinking.

PANELIST IDENTITY Busch and Jaeger (1990) found small, statistically significant differences in panelist judgments following group discussion in a modified Angoff standard setting for the National Teacher Examination in North Carolina. Specifically, public

school teachers tended to revise their judgments in the direction of teacher educator panelists. In the Angoff and Hofstee methods, Wayne et al. (2008) reported differences in recommended cut scores among panelists with different experiences (i.e., communication experts, program directors, postgraduate trainees, or hospital patients) for a medical communications exam. In contrast, Wyse (2018) did not detect significant differences in Angoff judgments across panelist professional roles for 14 medical imaging credentialing exams carried out over 15 years. He also explored potential impacts of examination development experience (e.g., prior standard setting, serving as an item writer or reviewer) on recommended cut scores and, again, did not find statistically significant effects. Cravens et al. (2013) found convergence in judgments across principals, supervisors, and teachers when implementing the Bookmark method for a principal evaluation tool.

Other studies present panelist effects related to content expertise, and for the most part suggest that content expertise does influence panelists' standard-setting judgments. An earlier study by Plake et al. (1994) did not find systematic differences in Angoff cut scores recommended by content specialists versus nonspecialists in various subjects on the College Basic Academic Examination. However, more recent studies in the medical field report substantive differences in judgments related to content knowledge. For example, Verhoeven et al. (2002) found that item writers recommended an Angoff cut score 12.4% higher than recent graduates for a medical school progress examination in the Netherlands. Evaluating expertise by asking panelists to complete test items, Verheggen et al. (2008) found that knowing the answers to undergraduate psychiatry items was associated with providing a higher Angoff cut score for most panelists. Margolis et al. (2016) also found that answering an item correctly was associated with higher Angoff judgments in medical licensure. Clauser et al. (2017), also in medical licensure, reported that panelists tended to judge items covering content that was unfamiliar to them as more difficult than items covering familiar content.

Research on impacts of standard-setting panelists' social identities is less common. Caines and Engelhard (2012) explored relationships among panelists' personal characteristics (i.e., race/ethnicity and gender), teaching contexts (i.e., school socioeconomic status and region), and recommended cut scores in modified Angoff standard settings for mathematics and English language arts assessments. They did not find race/ethnicity or gender to exert a substantial effect on panelists' recommended cut scores. However, they found significant impacts for school socioeconomic status (captured through upper, middle, and lower designations of students receiving free and reduced-price lunch in the district) and region (i.e., five geographic areas across the state). Specifically, teachers from schools in areas with low socioeconomic status recommended lower cut scores. Although region also appeared to influence teachers' judgments, it is difficult to interpret this substantively because the identity of the state in which the study was carried out is not provided. It is also possible that school socioeconomic status and region are correlated.

GROUP DISCUSSION Researchers have approached the topic of group discussion from several perspectives, examining both its content and its effects on recommended cut scores. In a research simulation of a modified Angoff standard setting, Novakovic (2008) found that mock panelists for a Teaching Exercise and Fitness certification test did not make statistically significant adjustments to their judgments following group discussion, a clear illustration of anchoring and adjustment bias (Kahneman, 2011).¹⁰ Panelists did self-report that the discussion influenced their decisions; this discussion appears to have been confirmatory for most participants. Some panelists did report experiencing peer pressure to modify their judgments to conform to the group. These panelists also reported that the introduction of empirical performance data into a subsequent round of judgments undermined their confidence. Clauser, Harik, et al. (2009), using a medical credentialing examination in the United Kingdom, showed substantial impact on the variability among panelists' estimates of borderline test takers after discussion of Round 1 results.

Several studies examine the content of group discussions in standard setting. In a study linking scores from two United Kingdom English language assessments of the Common European Frame of Reference (CEFR) using the basket standard-setting method¹¹ (Council of Europe, 2009; also see Kaftandjieva, 2010). Papageorgiou (2010) developed a nuanced account of group discussion by coding specific topics of conversation. Three coding categories relevant to panelist decision-making emerged, each of which contained several subtopics: the CEFR language proficiency standards, test takers, and the standard-setting panelists. When discussing the CEFR standards, panelists identified problems with the framework as well as difficulties they encountered throughout the standard-setting process, such as conceptualizing a borderline test taker rather than a test taker with clearly sufficient knowledge. The test-taker coding category included test-taker performance, the target test-taker definition provided by the CEFR performance levels, and explicit reference to borderline test takers who just passed one of the English language exams. Finally, the panelist code encompassed panelist confidence, references to specific test takers the panelists knew, panelist expectations for their students, and standard-setting feedback and consequences. Panelists most frequently discussed test-taker performance and target test-taker descriptions.

Deunk et al. (2014) brought discussion contents and effects together in their study of a Bookmark standard setting for reading and mathematics assessments in the Netherlands. The authors found that panelists did alter their bookmark placements following group discussion, although not in a consistent direction. Their content analysis of group discussion transcripts revealed five categories of panelist discussion: item content, school contexts, scores, the standard-setting meta task (i.e., the standard-setting process itself), and other topics not related to the standard-setting task. Surprisingly, some panelist groups did not discuss item content at all. In general, mathematics panelists spent more time discussing content than English panelists did. The degree to which the facilitator participated in, or guided, the discussion is not addressed.

POLICY AND SOCIOCULTURAL CONTEXTS Research on workshop contexts investigates the broader settings in which standard-setting panelists carry out their work. These include the policy landscape surrounding the standard setting and even the role of the assessment itself. Some of these investigations are quite specific; that is, carried out within a single standard-setting workshop. Others try to capture broader contexts using surveys. In Figure 12.2, we conceptualized the context as the larger container, political and social context. We provided a brief introduction to these factors as pertaining to the purpose of the test and intended use of test scores, which would reflect the cut scores to be set.

At a broad level, research sheds some light on aspects of the standard-setting design that may go unnoticed in individual standard-setting studies, for example, when designers favor a specific standard-setting method or use different descriptions of borderline test takers across contexts. In a survey of U.S. state assessment directors, Kannan (2014, 2016) reported that the Bookmark method appears to be the most widely used standard-setting method within K–12 state assessment. A majority of states also incorporate PLDs into their standard-setting processes. However, one state reported panelist difficulty in focusing solely on the PLDs without considering the policy implications of their task.

More specifically, the use of specific language can influence judgments given how terms are perceived, possibly because of the personal and professional implications of the standard-setting outcomes. In a survey of teachers, Burt and Stapleton (2010) examined how performance level labels can vary in evaluative connotations for teachers. Evaluative connotation refers to “the personal meaning individuals associate with a word” (Burt & Stapleton, 2010, p. 30). They found that different labels did in fact evoke different evaluations and judgments about mastery. For example, teachers judged the terms *Limited Knowledge*, *Basic*, and *Apprentice*—all used by different states to represent a generic *Basic* level—quite differently in terms of teacher evaluative connotation and mastery. Specifically, *Apprentice* was evaluated significantly more positively (i.e., reflecting a higher level of mastery) than both *Basic* and *Limited Knowledge*, and *Basic* was evaluated significantly more positively than *Limited Knowledge*.

More recently, O'Donnell (2020) surveyed teachers and parents regarding their perceptions of various performance-level labels used in middle school assessments across several states. Participants rated labels for encouragement (i.e., not potentially disparaging to students) and clarity and found differences in these ratings across different labels. For example, among the labels representing the lowest levels of performance, teachers found the labels *Standard Not Met* and *Did Not Pass* to be clearer than *Minimal* or *Below*.

Other research has focused specifically on standard setting outside the United States. Ferdous and Buckendahl (2013) collected data on panelists' thinking in an Angoff yes/no workshop for English and mathematics in a sub-Saharan Africa country. Panelists completed questionnaires at various junctures during the standard setting: after training, after Round 1 judgments, after feedback, and after Round 2 judgments. Panelists reported general comfort with the standard-setting process and confidence

in their judgments throughout the workshop. For Round 1 judgments, English and mathematics panelists reported thinking about students in their classes and PLDs and indicated generally that they did not provide judgments based on a desire to set high or low expectations for schools. Panelists were mixed in their perceptions of the utility of feedback and did not appear to make as much use of feedback as panelists from U.S. standard settings typically do. Panelists in this study also did not explicitly reference policy structures as influential in their decision-making.

Pitoniak and Yeld (2013) described challenges in the implementation of pilot and operational Angoff standard settings for National Benchmark Tests in South Africa. Panelists expressed concern throughout the process that any standards that were set might adversely affect Black students in historically underresourced schools. This led to a high level of anxiety throughout the process and may have negatively impacted panelist engagement in the task, as evidenced by repetitive questions about the process and some lack of attention to feedback provided between rounds of judgment. These observations led to revision of procedures between the pilot and operational studies. The authors suggested that pilot workshops are essential for standard setting in new contexts.

Summary and Connections to the Cognitive and Social Model of Standard Setting

The research we have reviewed on panelists' standard-setting experiences highlights the relevance of the cognitive and social components of the model of standard setting. Several studies present conclusions supporting the explicit consideration of panelist thinking and the components of the individual panelist cognitive processing submodel. Perhaps most important, we find that panelists' faithful and intentional application of the cognitive-judgmental tasks for variations of the Angoff and Bookmark methods—or any method—cannot be taken for granted. Some studies provide clear evidence of panelists' deviation from a method's intended cognitive-judgmental task and application of alternative decision-making strategies. Perhaps the clearest example of this is the use of algorithmic thinking to create rules for bookmark placements based on item proportions as a hypothesized explanation for the findings of Clauser et al. (2017), Davis-Becker et al. (2011), and Hein and Skaggs (2009). Other studies shed light on instances of potential deviation from a method's intended cognitive-judgmental task that may require further examination and explanation. For example, panelists in several studies (e.g., Dawber et al., 2002; Hein & Skaggs, 2010) reported thinking about their own students when making judgments, which raises concerns about the influence of the availability heuristic. Even in cases where panelists do not directly report difficulty carrying out prescribed standard-setting judgments, other data sources such as panelist surveys or interviews may indicate otherwise (e.g., Novakovic, 2008). These findings are consistent with Katz's (2019) framework of standard-setting panelist cognition for making probability judgments. Hein and Skaggs's (2009) finding that panelists used such strategies in the pursuit of "adequate justification" for their recommendations

(p. 219) is interesting to consider because it suggests that panelists may perceive alternative decision-making strategies as more defensible than faithful application of the cognitive-judgmental task prescribed by the standard-setting method. These findings are particularly noteworthy given the popularity of these methods.

Research also highlights the importance of considering the social components of standard setting represented in the group interactions submodel. Findings concerning the impact of group discussion on panelist decision-making are mixed, with some research suggesting that panelists report discussion to be more influential on their thinking than empirical data (e.g., Novakovic, 2008). Other research finds inconsistent changes in panelist judgments following group discussion (e.g., Deunk et al., 2014). The impacts of panelist professional identities is somewhat ambiguous, with earlier research (e.g., Busch & Jaeger, 1990; Wayne et al., 2008) finding differences in recommendations based on professional role that do not emerge in more recent studies (e.g., Cravens et al., 2013; Wyse, 2018).

It is also worth considering information that is not present within this review. The studies discussed here mostly concern probability judgments. We have found no research that explicitly examines panelist thinking or experiences in the other categories of the taxonomy. Thus, the ways in which panelists may struggle with or adapt matching, threshold detection, and value judgments in relation to the individual panelist cognitive processing submodel are largely unknown. There is also much that is unknown about the interplay of the components of the group interactions submodel. For example, we could locate only a single study that explicitly focused on panelist social identities (Caines & Engelhard, 2012). We could find no research that examines workshop facilitation. Research from judgment and decision-making and social psychology literatures supports these submodels. Further research is needed to examine their manifestation and influences in different standard-setting contexts. We propose what the beginning of such a research agenda might look like in the next section.

Standard-Setting Questions: A Proposed Research Agenda

Papageorgiou and Tannenbaum (2016) provided a comprehensive research agenda to support standard setting. They argued that “standard setting should be an integral part of the validity argument” (p. 109) rather than responding to allegations about its arbitrariness. Their research agenda was organized around substantive and procedural components of standard setting and borrowed extensively from Plake (2008), who proposed a research program that challenges standard-setting assumptions and practices.

We call attention to research questions derived from the cognitive and social model of standard setting proposed in this chapter. A comprehensive research agenda that addresses all parts and processes of the model would be massive. Therefore, the research agenda we propose here is only a starting point. Our recommendations focus on the influences of heuristics and biases on the cognitive-judgmental task and interpersonal interactions and highlight additional questions as recommended topics for future investigations.

Research Questions on Heuristic, Biases, and Other Influences

Within this chapter, we have integrated theories as to how particular heuristics and biases might be operationalized in various standard-setting methods. These hypotheses represent how we connected the research on judgment and decision-making with standard-setting processes and experience facilitating standard-setting activities. Within this section, we present a set of research questions that would help test these hypotheses and develop solutions to mitigate any negative impacts.

1. How does each heuristic or bias affect panelist thinking and decisions in different standard-setting methods? Which methods are most susceptible to the use of, or impact from, heuristics or biases? Specific examples of this research question include the following:
 - In which methods are panelists likely to use substitution?
 - When panelists are provided several sources of information to consider, do they rely on selected pieces more than others (e.g., use of availability and WYSIATI)?
 - Do panelists bring in additional memories or information while making their judgments?
 - How do panelists use the representativeness and recognition heuristics for efficient decision-making? How might these contribute to judgmental errors?
 - In what ways do anchoring and adjustment bias affect judgments in Rounds 2 or 3?
 - How pervasive are the effects of the affect heuristic in standard setting? Do they operate more or less for specific methods?
2. What are the sources of cognitive complexity and cognitive load for each standard-setting method?
3. Are particular groups of panelists more likely to use heuristics or biases (e.g., more or less experienced)?
4. What facilitation strategies, such as debiasing, can minimize the impact of heuristics and biases on panelist judgments?

Research Questions on Interpersonal Influences in Standard Setting

Although empirical evidence indicates that panelist judgments shift following group discussion, little research focuses on interpersonal interactions among panelists and between panelists and workshop facilitators (cf. Fitzpatrick, 1989). The group interaction submodel highlights likely influences regarding these interactions, as well as various aspects of panelists' social identities. The field could benefit from research concerning the following.

1. To what extent can standard setting be conceptualized as a group task and how might concepts from group work research (e.g., coordination, group member misbehavior; Barron, 2000; Kolb, 2016; Rico et al., 2008) appear in standard-setting workshops?

2. How do panelists' social identities (e.g., gender, race, profession) impact their participation in standard-setting workshops and influence the cut scores that they recommend?
3. To what extent is deep diversity (Harrison et al., 1998) achieved in standard-setting panels and how do personal beliefs and values impact panelist participation in standard-setting workshops and influence the cut scores they recommend?
4. What facilitator competencies and areas of expertise are most important for effective standard-setting facilitation?
5. How do facilitator actions influence panelists' engagement in a standard-setting method's cognitive-judgmental task?

Other General Research Questions

Once results from studies that address these research questions emerge, we can adjust standard-setting practices and or envision methodologies that will enable panelists to follow the cognitive-judgmental task with minimal negative influences. We also will be able to debate more general questions such as:

1. Should we continue to use traditional standard-setting methods given their susceptibility to heuristics and biases that undermine the fidelity of their cognitive-judgmental tasks? What judgmental methods could we use instead?
2. We have argued that the fidelity by which cognitive-judgmental tasks are implemented matters in making validity arguments about the suitability of a method and defensibility of performance standards generated by those methods. How much does method fidelity influence the acceptability of recommended cut scores?
3. How can we revise popular standard-setting methods to reduce cognitive complexity, cognitive load, and susceptibilities to the negative effects of judgmental heuristics and cognitive biases?
4. Many of the studies that we have discussed are single studies. Given the "replication crisis" in psychology, medicine, and other fields (see, for example, Camerer et al., 2018), where findings from studies are replicated at disturbingly low rates, can standard-setting findings be replicated? Under what conditions? To what degree do panelist effects, facilitator influences, and the policy and political environment influence standard-setting recommendations?
5. How can we triangulate standard-setting research findings using methods such as think-alouds, surveys, and interviews other than those used in the original studies?

CONCLUSION AND EVOLUTION OF STANDARD SETTING

We began this chapter with three claims. We proposed the following:

1. The cognitive-judgmental task (inherent in the standard-setting method) is a critical aspect of standard setting and should be central to the design, preparation, execution, and evaluation of the workshop.

2. In addition to the cognitive-judgmental task, there are other influences on panelist judgment and decision-making in standard setting. Some are design features and are desirable; some are not.
3. Standard-setting results should reflect a well-executed process that was led by experts who helped panelists make their judgments, avoiding undesirable influences.

To support these claims, we began at the center of the model and presented the taxonomy of cognitive-judgmental tasks. We then presented a list of positive and negative influences on panelist judgment and decision-making, including judgmental heuristics, cognitive biases, and other factors. We related these influences on the submodels in the overall model and provided suggestions as to how standard-setting facilitators can manage such influences during the standard-setting process and document validity evidence to this effect. Finally, we will highlight findings from standard-setting research that connects directly to these topics and make suggestions as to how these influences can be explored in future research.

State of the Art and Practice: Innovation and Inertia

Standard setting is evolving (e.g., Ferrara, 2022). At the same time, its practice remains essentially unchanged (Davis-Becker, 2022). What do we see for the near future of standard setting? We propose that recent innovations represent evolution in thinking about standard setting and some practice of standard setting (Ferrara, 2022; see also Luecht, 2020). The innovations included the following.

First are the innovations in standard-setting workshop design, including selecting a method with an appropriate cognitive-judgmental task, accounting for the implications in two proposed models of standard setting—the cognitive and social model we have proposed here and a psychometric model of standard setting proposed by Reckase (2023)—and remote standard setting workshops (Kanistra, 2022; Kollias 2018).

Second, there is an increasing influence of policy during standard setting, including conceptualizing standard setting as a lever to pursue policy aspirations for student achievement and for specification of minimal competency for new professionals, and Benchmarked Standard Setting approaches (Ferrara et al., 2021).

Third is the increasing prevalence in the use of principled assessment design, including writing PLDs as a first step in assessment design (Bejar et al., 2007; Huff et al., this volume, Luecht, 2020), selecting an appropriate method (Ferrara et al., 2023), articulating content standards and PLDs vertically (Cizek & Agger, 2012), pursuing item-PLD alignment (Ferrara, 2017), standards validation (Landl, 2019; Mattar et al., 2012), and embedded standard setting (Lewis & Cook, 2020).

Embedded Standard Setting as an Example of Innovation

Lewis and Cook (2020) have proposed embedded standard setting as an approach to standard setting that integrates identifying cut scores into a principled item development

process and as preferable to setting standards after operational administration. They describe the process and argue for its logic this way.

When item writers develop items that are designed to measure the KSAs associated with specific performance levels, then the second fundamental task that . . . standard-setting panels perform—making item-level judgments that operationalize the KSAs and allow them to be translated into appropriate cut scores—has also been fulfilled, again with more thought, transparency, and rigor than can typically occur during a relatively brief standard-setting workshop. (Lewis & Cook, 2020, p. 9)

And, more succinctly, “the practice of principled assessment design renders traditional standard-setting methodology redundant at best and contradictory at worst” (p. 8).

Specifically, this approach proposed by Lewis and Cook (2020) determines cut scores based on the targeted alignment to PLDs. Further, this approach seeks to minimize the impact on cut scores of items whose empirical scale locations do not align with their targeted PLDs (R. Cook, personal communication, May 11, 2022). Gaertner and Powers (2021) conducted a simulation study in which they compared 13 statistical and procedural methods for calculating cut scores to minimize bias and variability in cut scores. They found that the embedded standard setting weight method, which chooses a cut score to minimize distances between item-PLD misalignments, worked well when there was item-PLD mismatch (any amount: low, medium, or high) and some (a medium amount) of imbalance in the difficulty of the items. Of course, embedded standard setting can be completed only when IRT item parameters are available, usually after analysis of a field test or operational administration, when item scale locations are available, empirical item-PLD alignments can be determined, and cut scores can be applied.

Following the main premise in this chapter, embedded standard setting can be conceptualized as combining the cognitive judgments, which typically are part of standard setting, within the test development process and then conducting the empirical aspect of standard setting (i.e., calculation of recommended cut scores, evaluating validity evidence) after data are available. The goal of both traditional standard-setting processes and embedded standard setting is to maximize item–PLD alignment for a test form. The process can be applied to entire item banks, although cut scores should be based on selections of items that match test form blueprints.

A Recent Trend

Recently, standard-setting leaders and state departments of education have had difficulty in recruiting panelists for standard-setting workshops for state assessment programs. This difficulty has arisen even when recruiting for panels as small as 10 to 12 participants. This is quite different from the early 2000s, when it was relatively easy to fill out panels of 25 to 30 participants. Fifteen to 20 years ago, the then-new No Child Left Behind requirement that all states test students in Grades 3–8 and high school was controversial and performance standards were worrisome to states.

That newness probably stirred statewide and local interest—sometimes expressed as outrage—which motivated educators to participate in state assessment standard settings, to learn what standard setting was about, and, perhaps, to protect against what they considered unreasonably high performance standards. We personally observed school principals and district assistant superintendents attending standard-setting workshops and openly encouraging panelists to propose low cut scores. More recently, new state assessments and performance standards seem to be accepted as routine, which may reduce the willingness of local educational leaders to allow teachers to leave their classroom for two or more days to participate in standard-setting workshops.

Inertia in Standard-Setting Practice

Despite this list of innovations, inertia is another way to describe standard-setting practice over the past several decades. Although new methods are introduced every few years (see Table 12.1), many practitioners still rely on methods that were first described in the 1970s. This lack of change in practice is not unique to standard setting and can describe much of the larger testing industry. For example, Hambleton (1989) described the IRT work of Lord from the early 1950s, with some of the first large-scale applications occurring much later, in the early 1980s.

This inertia is somewhat understandable. First, many of the “new” methods are variations on presentation or judgmental guidance of existing methods. Second, the high-stakes nature of many testing programs discourages setting designers from implementing any operational changes. Third, many programs are seeking, or seeking to maintain, approval by an external organization (e.g., U.S. Department of Education peer review, third-party accreditation) and wish to use operational practices that have been previously approved by these organizations.

CLOSING

Whether you view standard setting as a rigorous integration of art and science or a necessary evil, it is a significant enterprise in educational and credentialing testing. The measurement community’s recognition of its importance has evolved from its first mention in the second edition of *Educational Measurement* in 1971 to its prominence in the early 21st century. It took only 50 years to get here. The importance of standard setting requires that we strive continuously to improve our methods, conceptualizations, and practices. We have attempted to meet that requirement in this chapter. We encourage our colleagues who are standard setters to continue their improvement efforts as well. And we hope that the cognitive and social model of standard setting, the taxonomy of cognitive-judgmental tasks, and the research agenda topics we have proposed here stimulate researchers to recognize the rich opportunities that standard setting provides for research topics.

ACKNOWLEDGMENTS

Steve Ferrara was the inspiration, driving force, and team leader for the development of this entire chapter. He was excited to share these ideas with the measurement community and inspire future debate and discussion around standard setting. We wrapped up the final draft of the chapter and were all excited to celebrate . . . except Steve, who was ready to work on the next research paper! Up until his passing, he was full of ideas, inspiration, and appreciation for this professional community. We are grateful for the opportunity to collaborate with Steve and hope everyone appreciates his ideas shared in this chapter.

The authors also wish to thank a number of our colleagues and friends who helped us during the development of this chapter. First, we want to thank the editors, Linda Cook and Mary Pitoniak, for the invitation to create this chapter and the support during the process. Second, we want to thank our reviewer/collaborators, Barbara Plake, Ellen Forte, and Jim Pellegrino, who provided guidance in shaping our work and communicating a clear message. Third, we want to thank other colleagues who contributed ideas and reviews based on their experience and expertise: Chad Buckendahl, Brian Clauser, Rob Cook, Jamie Dunleavy, Voula Kanistra, Charalambos (Harry) Kollias, Dan Lewis, Marianne Perie, and Mark Reckase. Finally, we want to thank the graphics team at HumRRO for their help in developing our visual model of standard setting: Constance Denning, Ryan McGrath, Haeley Phan, and Stephanie Means.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Psychological Association. (n.d.) Compliance. In *APA Dictionary of Psychology*. Retrieved January 27, 2025, from <https://dictionary.apa.org/compliance>
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). American Council on Education.
- Atash, N., Ferrara, S., & Bayless, D. (1994, June). *A-F-G-B: A new method for setting absolute standards* [Paper presentation]. Council of Chief State School Officers (CCSSO) Annual National Conference on Large-Scale Assessment, Albuquerque, NM, United States.
- Baldwin, P., Margolis, M., Clauser, B., Mee, J., & Winward, M. (2020). The choice of response probability in Bookmark standard setting: An experimental study. *Educational Measurement: Issues and Practice*, 39(1), 37–44.

- Baron, P., & Kannan, P. (2013, October 23–25). *Variance in standard setting judgments: Investigating cognitive sources* [Paper presentation]. Northeastern Educational Research Association Annual Meeting, Rocky Hill, CT, United States.
- Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *The Journal of the Learning Sciences*, 9(4), 403–436.
- Bazerman, M., & Moore, D. (2013). *Judgment in managerial decision making* (8th ed.). John Wiley & Sons.
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 31–63). JAM Press.
- Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45(2), 4–9.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137–172.
- Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement*, 21, 147–152.
- Bieri, J. (1955). Cognitive complexity-simplicity and predictive behavior. *The Journal of Abnormal and Social Psychology*, 51(2), 263–268. <https://doi.org/10.1037/h0043308>
- Bonito, J. A., & Sanders, R. E. (2011). The existential center of small groups: Member's conduct and interaction. *Small Group Research*, 42(3), 343–358.
- Brandon, P. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, 17(1), 59–88.
- Briggs, D. (2022). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge.
- Buckendahl, C. (Ed.). (2005). Qualitative inquiries of participants' experiences with standard setting [Special issue]. *Applied Measurement in Education*, 18(3).
- Buckendahl, C., & Davis-Becker, S. (2012). Setting passing standards for credentialing programs. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 485–502). Routledge.
- Buckendahl, C., Huynh, H., Siskind, T., & Saunders, J. (2005). A case study of vertically moderated standard setting for a state science assessment program. *Applied Measurement in Education*, 18(1), 83–98.
- Burnstein, E., & Vinokur, A. (1975). What a person thinks upon learning he has chosen differently from others: Nice evidence for the persuasive-arguments explanation of choice shifts. *Journal of Experimental Social Psychology*, 11(5), 412–426.
- Burt, W. M., & Stapleton, L. M. (2010). Connotative meanings of student performance labels used in standard setting. *Educational Measurement: Issues and Practice*, 29(4), 28–28.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27(2), 145–163.

- Caines, J., & Engelhard, G., (2012). How good is good enough? Educational standard setting and its effect on African American test takers. *Journal of Negro Education*, 81(3), 228–240.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature: Human Behavior*, 2(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed and new thoughts in second language testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 517–530). American Council on Education/Praeger.
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106.
- Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20–31.
- Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Lawrence Erlbaum Associates.
- Cizek, G. J. (Ed.). (2005). Vertically moderated standard setting [Special issue]. *Applied Measurement in Education*, 18(1).
- Cizek, G. J. (2006). Standard setting. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 225–260). Lawrence Erlbaum Associates.
- Cizek, G. J. (Ed.). (2012a). *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge.
- Cizek, G. J. (2012b). The forms and functions of evaluations of the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 467–484). Routledge.
- Cizek, G. J. (2012c). An introduction to contemporary standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). Routledge.
- Cizek, G. J. (2022, April 21–24). *Bad standard setting—and what to do about it* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.
- Cizek, G. J., & Agger, C. (2012). Vertically moderated standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 467–484). Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.
- Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 212–237). Routledge/Taylor & Francis Group.

- Clauser, J. C., Hambleton, R. K., & Baldwin, P. (2017). The effect of rating unfamiliar items on passing scores. *Educational and Psychological Measurement*, 77(6), 901–916.
- Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, 22, 1–21.
- Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, 46, 390–407.
- Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on the integration of performance data into Angoff judgments. *International Journal of Testing*, 13, 65–85.
- Cook, R., & Lewis, D. (2022, April 21–24). *Embedding standards to preserve intended score interpretation and use* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.
- Coraggio, J. T. (2008). *A Monte Carlo approach for exploring the generalizability of performance standards* [Unpublished doctoral dissertation]. University of South Florida.
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference (CEFR) for languages: Learning, teaching, assessment*. <https://rm.coe.int/1680667a2d>
- Cravens, X. C., Goldring, E. B., Porter, A. C., Polikoff, M. S., Murphy, J., & Elliott, S. N. (2013). Setting proficiency standards for school leadership assessment: An examination of cut score decision-making. *Educational Administration Quarterly*, 49(1), 124–160.
- Daly, J. A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, 19(4), 309–316.
- Davis-Becker, S. (2022, April 21–24). *How can we bring innovations into standard setting practice?* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Diego, CA, United States.
- Davis-Becker, S., Buckendahl, C., & Gerrow, J. (2011). Evaluating panelists' Bookmark standard setting judgments: The impact of random item ordering. *International Journal of Testing*, 11(1), 24–37.
- Dawber, T., Lewis, D. M., & Rogers, W. T. (2002, April 1–5). *The cognitive experience of bookmark standard setting panelists* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States.
- Deunk, M. I., van Kuijk, M. F., & Bosker, R. J. (2014). The effect of small group discussion on cutoff scores during standard setting. *Applied Measurement in Education*, 27(2), 77–97.

- Domaleski, C., D'Brot, J., Keng, L., Keglovits, R., & Neal, A.-M. (2018). *Establishing performance standards for school accountability systems*. Council of Chief State School Officers. <https://eric.ed.gov/?id=ED593538>
- Ebel, R. L. (1972). *Essentials of educational measurement*. Prentice-Hall.
- Egan, K. L., Schneider, C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–108). Routledge.
- Elliott, S. N., & Kettler, R. J. (2015). Item and test design considerations for students with special needs. In S. Lane, M. R. Raymond, and T. M. Haladyna (Eds.), *The handbook of test development* (2nd ed., pp. 374–391). Routledge.
- Environmental Protection Agency. (n.d.). *Drinking water regulations*. Accessed August 6, 2020, from <https://www.epa.gov/dwreginfo/drinking-water-regulations>
- Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). <https://congress.gov/114/plaws/publ95/PLAW-114publ95.pdf>
- Ferdous, A. A. (2005). *Item response theory-based approaches for computing cut scores from an Angoff-based standard setting study* [Unpublished doctoral dissertation]. University of Nebraska–Lincoln.
- Ferdous, A. A., & Buckendahl, C. W. (2013). Evaluating panelists' standard setting perceptions in a developing nation. *International Journal of Testing*, 13(1), 4–18.
- Ferdous, A. A., & Plake, B. S. (2005). The use of subsets of test questions in an Angoff standard setting method. *Educational and Psychological Measurement*, 65(2), 185–201.
- Ferdous, A. A., & Plake, B. S. (2007). Item selection strategy for reducing the number of items tested in an Angoff standard setting study. *Educational and Psychological Measurement*, 67(2), 193–206.
- Ferrara, S., Perie, M., & Johnson, E., (2002, April 2–4). *Setting performance standards: The item-descriptor matching procedure*. [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, United States.
- Ferrara, S., (2017, April 26–30). *Aligning knowledge and skill requirements in ALDs and item response demands* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Antonio, TX, United States.
- Ferrara, S. (2022, April 21–24). *Evolution in standard setting: We can expect changes in practice* [Coordinated paper session]. National Council on Measurement in Education Annual Meeting.
- Ferrara, S., & Lewis, D. M. (2012). The item-descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 255–282). Routledge.
- Ferrara, S., Lewis, D., & D'Brot, J. (2021). Setting benchmarked performance standards: A method, procedures, and empirical results. *Journal of Applied Testing Technology*, 22(1), 12–24. <https://www.testpublishers.org/journal-of-applied-testing-technology>
- Ferrara, S., Roussos, L., & Qin, Q. (2023). An analytical approach to argumentation for validity argument technical reports [Manuscript submitted for publication]. Cognia.

- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117–140.
- Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. *Review of Educational Research*, 59(3), 315–328.
- Gaertner, M., & Powers, S. (2021, May 18–June 11). *How to win at standard setting with imperfect data* [Paper presentation]. National Council on Measurement in Education Annual Meeting.
- Galton, F. (1907). Vox populi. *Nature*, 75(1949), 450–451.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4), 650–669.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704.
- Giraud, G., Impara, J., & Plake, B. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education*, 18(3), 223–232.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–521.
- Glaser, R., & Klaus, D. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in system development* (pp. 421–427). Holt, Rinehart & Winston.
- Green, D. R., Trimble, C. S., & Lewis, D. M. (2005). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, 22(1), 22–32.
- Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, 21(1), 16–22.
- Haertel, E. H., Beimers, J., & Miles, J. (2012). The briefing book method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods and innovations* (2nd ed., pp. 283–300). Routledge.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). American Council on Education.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). American Council on Education/Praeger.
- Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). Routledge.
- Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41–55.
- Hansen, M. A., Lyon, S. R., Heh, P., & Zigmond, N. (2013). Comparing panelists' understanding of standard setting across multiple levels of an alternate science assessment. *Applied Measurement in Education*, 26(4), 298–318.

- Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *The Academy of Management Journal*, 41(1), 96–107.
- Harvey, A. L., & Way, W. D. (1999, April 19–23). *A comparison of web-based standard setting and monitored standard setting* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Montreal, QC, Canada.
- Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world: The psychology of judgment and decision making* (2nd ed.). Sage Publications.
- Hayes, B., & Heit, E. (2013). How similar are recognition memory and inductive reasoning? *Memory & Cognition*, 41(5), 781–795.
- Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the Bookmark method. *Applied Measurement in Education*, 22(3), 207–228.
- Hein, S. F., & Skaggs, G. E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experiences during standard setting. *Educational Measurement: Issues and Practice*, 29(2), 36–44.
- Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 109–127). Jossey-Bass.
- Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for K–12 test-based accountability. *Measurement: Interdisciplinary Research and Perspectives*, 8(2–3), 130–144.
- Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59(6), 885–897.
- Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34(4), 353–366.
- Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis*, 4(4), 461–475.
- Jaeger, R. M. (1989). Certification of student competence. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Prentice-Hall.
- Jaeger, R. M. (1995a). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15–40.
- Jaeger, R. M. (1995b). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, 14(4), 16–20.
- Jaeger, R. M., & Mills, C. N. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 313–338). Lawrence Erlbaum Associates.

- Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six methods with an application to tests of reading in EFL*. CITO.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark, Hachette Book Group.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty. heuristics and biases*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809477>
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425–461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Lawrence Erlbaum Associates.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Kanistra, V. (2022, April 8–14). *Turning the page on benchmarking and standard setting studies*. [Paper presentation]. National Council on Measurement in Education Annual Meeting.
- Kannan, P. (2014). *Content and performance standard articulation practices across the states: Report summarizing the results from a survey of the state departments of education* (ETS Research Memorandum RM-14-09). ETS.
- Kannan, P. (2016). *Vertical articulation of cut scores across grades: Current practices and methodological implications in the light of the next generation of K–12 assessments* (Research Report RR-16-29). ETS. <http://dx.doi.org/10.1002/ets2.12115>
- Kannan, P., Sgammato, A., Tannenbaum, R. J., & Katz, I. R. (2015). Evaluating the consistency of Angoff-based cut-scores using subsets of items within a generalizability theory framework. *Applied Measurement in Education*, 28(3), 169–186.
- Kannan, P., Sgammato, A., & Tannenbaum, R. (2015). Evaluating the operational feasibility of using subsets of items to recommend minimal competency cut scores. *Applied Measurement in Education*, 28(4), 292–307.
- Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4–12.
- Katz, I. R. (2019). *Standard setting panelist cognition: A framework and implications for practice* (Research Memorandum RM-12-12). ETS.
- Katz, I. R., & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology*, 15(1), 1–17.
- Katz, I. R., Tannenbaum, R. J., & Kannan, P. (2009). Virtual standard setting. *CLEAR Exam Review*, 20, 19–27.

- Kester, L., & Kirschner P. A. (2012). Cognitive tasks and learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 619–622). Springer. https://doi.org/10.1007/978-1-4419-1428-6_225.
- Kingston, N. M., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 219–248). Lawrence Erlbaum.
- Kingston, N., & Tiemann, G. (2012). Setting performance standards on complex assessments. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201–224). Routledge.
- Kolb, J. A. (2016). How not to work and play with others: An updated taxonomy of dysfunctional individual behaviors in groups. *Performance and Improvement*, 55(2), 12–19.
- Kollias, C. (2018, May 24–27). *Virtual standard setting in language testing. Exploring the use of synchronous audio and audio-visual environments* [Paper presentation]. European Association for Language Testing and Assessment 15th Annual Conference, Bochum, Germany.
- Landl, E. (2019, February 28). *Re-envisioning performance standards validation*. Center for Assessment. <https://www.nciea.org/blog/standard-setting/re-envisioning-performance-standards-validation>
- Larrick, R. P. (2007). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Handbook of judgment and decision-making* (pp. 316–337). Blackwell.
- Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement Issues and Practice*, 39(1), 8–21.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavior anchoring* [Symposium]. Council of Chief State School Officers (CCSSO) Annual National Conference on Large Scale Assessment, Phoenix, AZ, United States.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, M. (2012). The Bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 249–272). Routledge.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press.
- Lindquist, E. F. (1951). *Educational measurement*. American Council on Education.
- Linn, R. L. (1989). *Educational measurement* (3rd ed.). American Council on Education.
- Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. ETS.
- Loomis, S., & Bourque, M. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 175–218). Lawrence Erlbaum Associates.

- Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, 47(6), 1231–1243.
- Luecht, R. M. (2020). Generating performance-level descriptors under a principled assessment design paradigm: An example for assessments under the Next Generation Science Standards. *Educational Measurement: Issues and Practice*, 39(4), 105–115.
- Margolis, M. J., & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice*, 33(1), 15–22.
- Margolis, M. J., Mee, J., Clauser, B. E., Winward, M., & Clauser, J. C. (2016). Effect of content knowledge on Angoff-style standard setting judgments. *Educational Measurement: Issues and Practice*, 35(1), 29–37.
- Mattar, J., Hambleton, R. K., Copella, J., & Finger, M. (2012). Reviewing or revalidating performance standards on credentialing examinations. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 399–412). Routledge.
- McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Mills, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher*, 42(2), 78–88. <https://doi.org/10.3102/0013189X12470855>
- Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice*, 32(3), 27–35.
- Merriam-Webster. (n.d.). Value judgment. In *Merriam-Webster dictionary*. Retrieved June 29, 2020, from <https://www.merriam-webster.com/dictionary/value%20judgment>
- Morgan, D. L. (2004, June 21–23). *The performance profile method (PPM): A unique standard setting method as applied to a unique population* [Paper presentation]. Council of Chief State School Officers (CCSSO) Annual National Conference on Large-Scale Assessment, Boston, MA, United States.
- Morgan, D. L., & Michaelides, M. P. (2005). *Setting cut scores for college placement* (Research Report No. 2005–9). College Board.
- Morgan, D. L., & Perie, M. (2004). *Setting standards in education: Choosing the best method for your assessment and population* [Unpublished manuscript]. ETS.
- Murphy, G. L. (2004). *The big book of concepts*. MIT Press.
- National Commission for Certifying Agencies. (2004). *Standards for the accreditation of certification programs*. Institute for Credentialing Excellence.
- National Commission for Certifying Agencies. (2014). *Standards for the accreditation of certification programs*. Institute for Credentialing Excellence.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14(1), 3–19.
- Nichols, P., Twing, J., Mueller, C., & O'Malley, K. (2010). Standard-setting methods as measurement process. *Educational Measurement: Issues and Practice*, 29(1), 14–24.

- Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Lawrence Erlbaum Associates.
- No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2008).
- Novakovic, N. (2008). Thinking on the edge: The influence of discussion and statistical data on awarders' perceptions of borderline candidates in an Angoff awarding meeting. *International Journal of Training Research*, 6(1), 74–102.
- O'Donnell, F. (2020). *What's in a label? Unpacking the meaning of achievement labels from tests* [Doctoral dissertation, University of Massachusetts Amherst]. <https://doi.org/10.7275/15558737>
- O'Malley, K., Keng, L., & Miles, J. (2012). From Z to A: Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards: Foundation, methods and innovations* (2nd ed., pp. 301–322). Routledge.
- Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing*, 27(2), 261–282.
- Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, 13(2), 109–123.
- Pellegrino, J., Jones, L., & Mitchell, K. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress* (Report of the Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment, Commission on Behavioral and Social Sciences and Education, National Research Council). National Academies Press.
- Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15–29.
- Perie, M., & Thurlow, M. (2012). Setting achievement standards on assessments for students with disabilities. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 347–378). Routledge.
- Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 323–346). Routledge.
- Pitoniak, M., & Yeld, N. (2013). Standard setting lessons learned in the south African context: Implications for international implementation. *International Journal of Testing*, 13(1), 19–31.
- Plake, B. S. (2008). Standard setters: Stand up and take a stand. *Educational Measurement: Issues and Practice*, 27(1), 3–9.
- Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181–200). Routledge.
- Plake, B. S., & Hambleton, R. K. (2001). The analytical judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 283–312). Lawrence Erlbaum Associates.

- Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement*, 57(3), 400–411.
- Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement*, 31(4), 339–347. <http://www.jstor.org/stable/1435397>
- Plous, S. (1993). *The psychology of judgment and decision-making*. McGraw–Hill.
- Psychology Wiki. (n.d.). Value judgment. In *Psychology Wiki*. Retrieved June 29, 2020, from https://psychology.fandom.com/wiki/Value_judgment
- Raymond, M. R., & Reid, J. R. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Lawrence Erlbaum Associates.
- Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–173). Lawrence Erlbaum Associates.
- Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard-setting methods. *Educational Measurement: Issues and Practice*, 25(2), 4–18.
- Reckase, M. D. (2023). *The psychometrics of standard setting: Connecting policy and test scores*. CRC Press.
- Reckase, M. R., & Chen, J. (2012). The role, format, and impact of feedback to standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 149–164). Routledge.
- Rico, R., Sánchez-Manzanares, M., Gil, F., & Gibson, C. (2008). Team implicit coordination processes: A team knowledge-based approach. *The Academy of Management Review*, 33(1), 163–184.
- Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). University of Minnesota, National Center on Educational Outcomes.
- Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 129–152). Cambridge University Press.
- Rottenstreich, Y., & Kivetz, R. (2006). On decision-making without likelihood judgment. *Organizational Behavior and Human Decision Processes*, 101(1), 74–88.
- Schnipke, D., & Becker, K. (2007). Making the test development process more efficient using web-based virtual meetings. *CLEAR Exam Review*, 18(1), 13–17.
- Shaw, R. (2007). Technology and testing. *CLEAR Exam Review*, 18(1), 7–9.
- Shepard, L. (1978). Special issue on standard setting. *Journal of Educational Measurement*, 15(4).
- Shepard, L. H., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement. A report of the National Academy of Education*

- Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels.* Stanford University, National Academy of Education.
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2), 129–138.
- Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Review*, 15(1), 21–25.
- Skaggs, G., Hein, S. F., & Awuor, R. (2007). Setting passing scores on passage-based tests: A comparison of traditional and single-passage bookmark methods. *Applied Measurement in Education*, 20(4), 405–426.
- Skorupski, W. (2012). Understanding the cognitive processes of standard setting panelists. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 135–148). Routledge.
- Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, 18(3), 233–256.
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Why study risk perception? *Risk Analysis*, 2(2), 83–93.
- Smarter Balanced Assessment Consortium. (2015). *Achievement level setting final report*. <https://portal.smarterbalanced.org/library/en/achievement-level-setting-final-report-no-appendix.pdf>
- Smarter Balanced Assessment Consortium. (2017). *Smarter Balanced cut score validation final report*. <https://portal.smarterbalanced.org/library/en/smarter-balanced-cut-score-validation-final-report.pdf>
- Speekenbrink, M., & Shanks, D. R. (2013). Decision-making. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 682–703). Oxford University Press.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.
- Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Lawrence Erlbaum Associates.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Tannenbaum, R. J., & Kannan, P. (2015) Consistency of Angoff-based standard setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment*, 20(1), 1, 66–78.
- Thompson, N. (2022, April 22). *What is the Beuk compromise?* ASC. <https://assess.com/beuk-compromise/>
- Tversky, A., & Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453–458.
- Verheggen, M. M., Muijtens, A. M. M., Van Os, J., & Schuwirth, L. W. T. (2008). Is an Angoff standard an indication of minimal competence of examinees or of judges? *Advances in Health Sciences Education*, 13, 203–211.

- Verhoeven, B. H., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education*, 36(9), 860–867.
- Wayne, D. B., Cohen, E., Makoul, G., & McGaghie, W. C. (2008). The impact of judge selection on standard setting for a patient survey of physician communication skills. *Academic Medicine*, 83(10), 17–20.
- Wills, A. J. (2013). Models of categorization. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 346–357). Oxford University Press.
- Wyse, A. E. (2011). The similarity of bookmark cut scores with different response probability values. *Educational and Psychological Measurement*, 71(6), 963–985.
- Wyse, A. E. (2018). Examining how professional roles and test development experiences impact Angoff ratings. *Applied Measurement in Education*, 31(4), 324–334.
- Zieky, M. J. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 19–51). Lawrence Erlbaum Associates.
- Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on basic skills assessment tests*. ETS.
- Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cut scores: A manual for setting standards of performance on educational and occupational tests*. ETS.

NOTES

1. In standard setting, the term “Borderline Examinee” is more commonly used. However, we have used “Borderline Test Taker” to be consistent with the terminology used throughout the volume.
2. For example, as in the “Tom W’s Specialty” experiment (Kahneman, 2011, pp. 146–148): Tom is described to resemble stereotypes of nerdy engineers and computer scientists (not to leave out psychometricians and statisticians); experimental subjects select those graduate specializations as most likely for Tom even though higher probability guesses—that is, specializations with higher enrollments and base rates—are humanities and education.
3. Similarities in evidence, with disregard for the veracity of the evidence.
4. We make decisions based only on those ideas that we can activate in working memory, which can lead to biased and suboptimal judgments.
5. Replacing a difficult question or cognitive task with a simpler one and then applying the judgment to the original question or task.
6. Voting and Supreme Court decisions are processes in which consensus (i.e., an agreement held by all or most) is sought, though usually not achieved. Juries arrive at verdicts by consensus, though television shows and movies suggest that it often is a wearing-down process rather than true agreement. We recognize that Galton’s (1907) observation was about accuracy. In our view, reasonableness

and acceptability of the group median cut score is analogous conceptually to accuracy.

7. Although there is no way to judge cut scores as correct or incorrect, “reasonable” means the cut scores (or resulting impact) would be judged as fair or appropriate by informed stakeholder groups.
8. Persuasive argument theory refers here to effects on panelists when they receive insights and other information from other panelists, not that panelists are explicitly trying to persuade other panelists.
9. In the traditional Bookmark method, items are ordered by difficulty, disregarding passage item sets. The single-passage Bookmark method is a variation applied to reading assessments with reading passages and intact item sets. Separate OIBs are created for each passage and item set in an effort to reduce the cognitive complexity of the task (Skaggs et al., 2007).
10. People estimate or accept an initial value, then make inadequate adjustments toward accuracy.
11. This is a standard-setting method, similar to Body of Work and ID Matching, in which panelists put test items in “baskets” that correspond to one of the CEFR achievement levels relevant to the test in question.

Reporting Scores and Other Results

April L. Zenisky

University of Massachusetts Amherst

Francis O'Donnell

NBME

Ronald K. Hambleton

University of Massachusetts Amherst

When most people who are not test developers think about tests and measurements, it is not models or theories that typically come to mind. It is test scores and how they have been presented and explained that is memorable and often consequential. Basic questions often asked, such as “Did you pass?,” “How’d you do?,” and “What does your score mean?,” speak to the impact that tests and test results have for stakeholders and where these users’ primary interests reside. Tests, by and large, fulfill a specific purpose for specific users, and there are actions and choices to be made on the basis of the scores, such as advancement down a career path, needing to take a review course or sign up for tutoring, or applying—or not—for admission to certain institutions. For these reasons, it is critical that test takers and other score users understand what the results mean relative to whatever actions may be available given the validated interpretation(s) and use(s) of the results.

Reporting test results is thus perhaps the single most important point of interaction between the testing agency and users of test scores, and, as the old saying goes, “You never get a second chance to make a first impression.” A well-designed and carefully crafted report of test results should support stakeholders by first communicating information about performance and, second, providing evidence to support whatever action(s) follow from that report of performance. In contrast, a poorly designed or confusing report does not lend itself to accomplishing an informational or actionable purpose and, in discouraging understanding of results, also can foster skepticism and/or distrust in a test.

GOALS FOR THE CHAPTER AND ADVANCE ORGANIZER

In this chapter, the intent is to provide readers with a broad theoretical context for understanding reporting and reporting choices and to draw on the growing body of research on reporting test results that is relevant to practical problems. To be clear, the approach here is less on prescriptive findings from prior studies and more on offering general trends where appropriate, tempered by the guidance for report development teams to engage with users whenever possible, to the extent possible.

The chapter begins with an overview of frames and perspectives for reporting, acknowledging that reports come in various sizes, shapes, and forms, and there are several conceptual approaches to reporting that are reflective of specific stakeholders’ interests and needs. Reporting efforts and report design approaches must be explicitly and directly tied to validity, meaning they must—first and foremost—support the intended interpretations and uses of test results. From there, the focus will turn to defining reporting in terms of several key dimensions, including the unit of analysis for reporting results, assessment contexts, report contents, mechanisms for reporting, and the context of reporting efforts within broader reporting systems.

Then, a seven-step model for reporting based on a synthesis of the work done by Hambleton and Zenisky (2013) is introduced to present the report development process

in light of the literature in this area, and an updated checklist for designing reports is provided. Topics of interest in this section include some reflection on the interdisciplinary nature of reporting and what can be learned from other fields, the need for documentation such as user guides alongside reports, and the need for ongoing research on reporting. From there, the chapter concludes with attention given to challenges and opportunities in reporting, such as next-generation assessments, subscore reporting, technology-based reporting, advances in data visualization, and reporting in a formative context.

CONTEXT FOR REPORTING AND RESEARCH

Over the years, I have been struck by the contradiction between the efforts and successes in producing sound technical assessments, drawing samples, administering the assessments, and analyzing the assessment data and the effort and success in disseminating the assessment results. (Hambleton, 2002, p. 193)¹

The body of research on educational and psychological measurement largely came into being in the 20th century. During that time, most of the more well-known tests were, as an unofficial rule, both norm referenced and multiple choice, and the focus of the measurement field was largely on advancing the technical, behind-the-scenes statistics and procedures to develop tests and validate proposed uses (see Clauser, 2019, for a history of classical test theory; see also Clauser et al., this volume, for a history of educational measurement). Communication with users has not always been prioritized as a matter of theory *or* practice, nor was the compilation of evidence establishing the potential score interpretations as reliable and valid themselves, separate from scores.

In reflecting on the history of *Educational Measurement* as a volume chronicling key elements of this field, it is worth noting that results reporting only gained its own chapter in this edition. This is not to say that reporting has been deliberately or consciously neglected in the editions curated by Lindquist (1951), Thorndike (1971), Linn (1989), and Brennan (2006). It has been discussed briefly in prior editions: Indeed, for example, in the 1951 edition, Mosier's chapter on "Batteries and Profiles" (Mosier, 1951) provided some guidance on the elementary principles of profile construction as

a graphic representation of a set of test scores for a single individual in which the tests are represented by ordinates spaced along the horizontal base line and the magnitude of each score is represented by plotting the point at the appropriate height on that ordinate. (p. 795)

In the 2006 edition, Cohen and Wollack noted that reporting is "clearly essential to test validity," but at the time it was "only just beginning to receive rigorous attention" (p. 382). Now, the time has come.

Reporting test results is a topic that cuts across psychometrics and many additional fields, including communications and marketing, graphic design, cognitive science, and information processing. In the same way that tests and testing practices have evolved, and stakeholder interest in and use of reports has changed along with how stakeholders

consume and use data more generally (Zenisky & Hambleton, 2012b), the way this profession has handled reporting has necessarily changed and is still evolving. No longer are reports simply data dumps with the goal to get as much information as possible onto a single page, and the literature on reporting is now much more substantial than even in the early 2000s.

A number of historical events and movements have also contributed to greater awareness of the impact of results reports. Reflect for a moment on several of the tests, testing programs, and assessment-related initiatives that have appeared over time at the forefront of the public consciousness involving tests: the introduction of Army Alphas and Army Betas for military selection in the early 20th century (Waters, 1997); the rise of the SAT (formerly Scholastic Aptitude Test) with the objective to level the playing field for college admissions (and waves of pushback to the use of standardized tests for that purpose) (Lemann, 1999); the publication of *A Nation at Risk* during the Cold War (National Commission on Excellence in Education, 1983; the standards-based revolution and accompanying transition from norm- to criterion-referenced testing (Hamilton et al., 2009); the No Child Left Behind Act of 2002 in the United States and the accountability imperative of testing affecting most schools and its subsequent reauthorizations (Linn et al., 2002); the rising interest in national and international comparisons afforded by assessments such as the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS) (Fischman et al., 2019); and the evolving “opt-out” movement in many places (Kirylo, 2018). What is undeniable through reflection on these (selected) watershed moments in the history of assessment is the social impact of test results and how communication of results and the context for those results can both reflect and refract historical shifts in society. However, it is equally important to note that the quality of the reporting of results across these initiatives is variable: Typically, the data that are gathered lend themselves more easily to summative statements about proficiency on a relatively general or global level, and many reporting efforts falter when it comes to providing actionable guidance for intended users of such data, particularly when it comes to instructional uses.

Indeed, at present, stakeholders are often more skeptical consumers of data and seek additional information whenever possible, leading to something of a tension between what stakeholders want and what interpretations the psychometric models strictly support. In this way, testing agencies must thread a needle of reporting transparency, quality, and quantity that speaks to the very fundamentals of test development and validation where, it is our belief, results reporting should be among the very first considerations. Results reports are often the final and most public-facing aspects of any assessment system. But, rather than being left to something that is handled late in the development process, within the context of a principled approach to assessment design, reporting must be prioritized as part of assessment development and validity considerations to inform all the decisions that follow.

What Is a Results Report?

The topic of reporting begins easily enough with a paper document, often a single page, printed double-sided, that provides test results for one individual. Figure 13.1 provides an example based on a fictitious K–12 testing program. This kind of report is broadly representative of what an individual might receive after taking a college admissions test, a standardized educational assessment, a credentialing test, or a psychological test battery. The information communicated on the individual score report in each of those areas can be tailored to the recipient, but the commonality is an accounting of the performance of a single individual on a test taken. The actual results included can be represented in a wide range of ways, such as with numbers, text/labels, and graphics, and the primary aim of this document is to be informative at that individual level (“Here’s how you performed”). In some testing contexts, paper reports have increasingly given way to electronic versions of the same information. These are sometimes accessed via email or, increasingly, a secure login to a reporting portal of some kind, but the reported information remains generally the same for individuals regardless of report delivery mode. Such reports can also be formatted for printing if that is an anticipated action users will take.

For some individual reports, it should be noted that the report not only serves the function of reporting status, but also offers prediction and/or guidance for next steps. This is perhaps most common in educational assessment, but is also found in reports

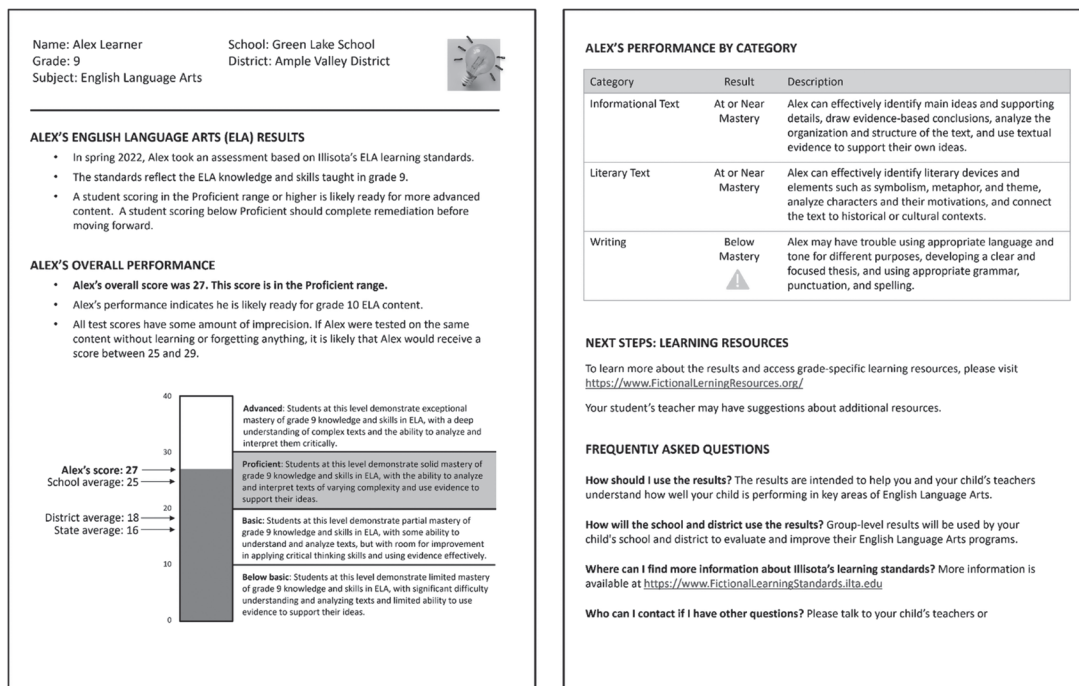


FIGURE 13.1
Simplified Sample of a K–12 Results Report

for other contexts where test users need help in identifying areas for improvement (for example, failing candidates in credentialing). This element of reporting can use subscore reporting techniques to identify areas of relative strength and weakness; in addition, rule-based/logic triggers (or, increasingly, artificial intelligence) can be implemented to evaluate response data and specify actionable next steps on a report. An example of this is a reading test that, as part of the reported results, provides a Lexile level to an individual and then suggests texts by name that are level appropriate (see Figure 13.2), or a mathematics assessment that reports relative weaknesses and structures reporting of such skills with links to specific online math lessons. Such connections, when included in reports, offer direct actions to the test taker and/or other users that take some of the guesswork out of what comes next, which is especially valuable when the report targets a test taker or families, who may well benefit from direct scaffolding in the area of “next steps.”

Group reports are also an essential part of the reporting landscape, and the format and audiences for those are more varied than for individual reports. It is when results for groups of test takers start being aggregated that the extent of the *who* and *what* of group reporting comes into view, and the idea of what a “report” is starts to expand exponentially. Consider the *who* of reporting for a moment: In an educational/K–12 setting, a group could, for example, consist of:

- a subset of students in a classroom, selected purposefully by a teacher or instructional administrator;
- all students in a classroom;
- selected students within a grade level grouped by some socioeconomic or demographic quality;
- all students in a grade level within one school in one district;
- all students in a grade level in multiple schools within a district;
- all students in a grade level within a state or territory; and/or
- all students in a grade level within a country.

These various groupings are not of equal interest to all users of assessment results. What a classroom teacher will focus on in a group report is necessarily quite different from what a school or district administrator will look at, which is different again from what

Your Reading Level: 1340L Current Reading Range: 1240L – 1390L		The best way to improve your reading level is to read books in your range. Consider the suggestions below.
Book	Author	Level
Hope For Animals and Their World	Jane Goodall	1240L
Stories Well Told: Science Fiction	Valerie Bodden	1340L
Core Four	Phil Pepe	1390L

FIGURE 13.2
Sample Report Section With Lexile-Based Book Suggestions

administrators in a state or provincial office of education consider most important. Although the idea of group reports is simple in that a group report is most basically some kind of aggregation of individual performance results, it broadly encompasses many different groups of individuals, data about their performance, and levels of communicating results.

Similar kinds of conceptualizations of what is a “group” for reporting can be applied to other testing settings. For example, in credentialing, a group could be all test takers testing at a certain test center within a defined testing window, all test takers from a certain institution or training school within a year, or all test takers grouped by a socio-economic or demographic quality within a year—these are dependent, again, on the audience and the planned use of the data.

Because of the many possible group selections for aggregate reports, group reporting is increasingly occurring (or at least originating) as a function of database queries rather than static report documents. To a point, dependent on the user group and the intended use(s) of group-level data, the process of accessing aggregate test results involves using interactive, web-based tools that tend to reflect a continuum of analysis (Zenisky & Hambleton, 2012b). This continuum spans from tools that provide results that are purely descriptive in nature to those that function more akin to a statistical analysis package to carry out original analyses, including significance testing. Examples of the former include the data tools for many U.S. states, such as Massachusetts (<http://profiles.doe.mass.edu/statereport/mcas.aspx>), Florida (<https://edstats.fldoe.org/SASPortal/main.do>), and Texas (<http://texasassessment.com/administrators/>). The International Data Explorer tool is the premiere example of the latter class of online assessment results tools because it is a “one-stop shop” for access to results from PISA, PIRLS, TIMSS, and the Programme for the International Assessment of Adult Competencies (<https://nces.ed.gov/surveys/international/ide/>). This portal, which mirrors that of the NAEP Data Explorer (<https://nces.ed.gov/nationsreportcard/data/>), permits users to run simple or complex descriptive results and carry out analyses of performance data with respect to innumerable additional variables gathered through participant and educator surveys. Results can be run to produce tables as well as highly customizable graphics.

Conceptualizing Reports and Reporting

Know the communicative purpose of the display and do not try to do too much.
(Wainer et al., 1999, p. 304)

A critical aspect of reporting that must be raised at the outset of this discussion concerns the idea of *effectiveness*. Much of the research and discussion that focuses on results reporting aims to support the development of “good” reports, but the quality of a report is a concept that deserves some reflection. What is an effective report? Ryan (2006) established a report itself as a form of communication, and a starting place to answer this question can be traced back to several key advances in the results reporting literature. First, the work of Wainer in focusing on visual displays of quantitative data (e.g., Wainer, 1992, 1997a, 2005, 2009; Wainer et al., 1999) is especially

instructive in that it draws across disciplines to formulate broad principles that apply to individual and group reports alike, such as creating visuals that are high in information and low in adornment, labeling clearly and fully, and using spacing to aid perception.

This line of research advanced by Wainer and others calls on those responsible for developing reports to think critically about the intended story of each data display as well as the nature of the data being presented. There are numerous examples in Wainer's writings of cases where graphics (assessment related and otherwise) were revised to show a very different story, one that was obscured by the original data presentation. Building on Wainer's work on graphical presentation and its considerable implications for results reporting of assessment data, then, effective—in part—draws on adhering to basic principles of communication and cognitive processing to create reports that are purposefully crafted to communicate specific information.

The foundational work on quality reporting done by Jaeger (Jaeger, 2003; Jaeger et al., 1993) likewise laid the groundwork for differentiation of reporting by audiences. By referring to "NAEP's audiences" (in the context of the NAEP) and reframing school report cards with the intended recipient in mind, reporting has effectively had to shift from one size fits all to a model that acknowledges and responds to the informational needs and interests of users of test data. This is a perspective that has rightly permeated current thinking about reporting and provides a conceptual basis for reporting efforts that are informed by research with various articulated user groups.

Figure 13.3 provides an updated perspective on Jaeger's (2003) nine tables offering a road map for audience-specific research activities for report development. The premise of Jaeger's (2003) work was that there are some key questions that should be asked of each stakeholder group, to learn not only what is of interest to them, but also how such information could and should be provided to them. And, with each stakeholder group, there are different methodologies available to report developers for the purpose of investigating those key questions, and those methods, naturally, provide quantitatively and qualitatively different data. Figure 13.3 links the main considerations in the design of research on reporting: For agencies that engage in these questions, it may be helpful to frame the work to be done relative to these considerations and build out the work relative to the priorities and available resources of a testing program.

Indeed, Jaeger's (2003) contributions, including the differentiation of users (and the accompanying need to probe and understand reporting interests of different groups), provide a direct line to another key advance in conceptualizing quality reports, which is the presentation of general models for report development and evaluation (Hambleton & Zenisky, 2013; Zapata-Rivera, 2011; Zenisky & Hambleton, 2012a, 2015). With regard to model-based approaches to report development, the focus on audiences and users as something to be articulated and respected propagated a shift in the process of report development, where Hambleton and others suggested that reports should evolve through audience-specific research and development, and reporting aims are discussed and user input is solicited at multiple points, to inform report development from conceptualization to implementation and maintenance. This advance helps to

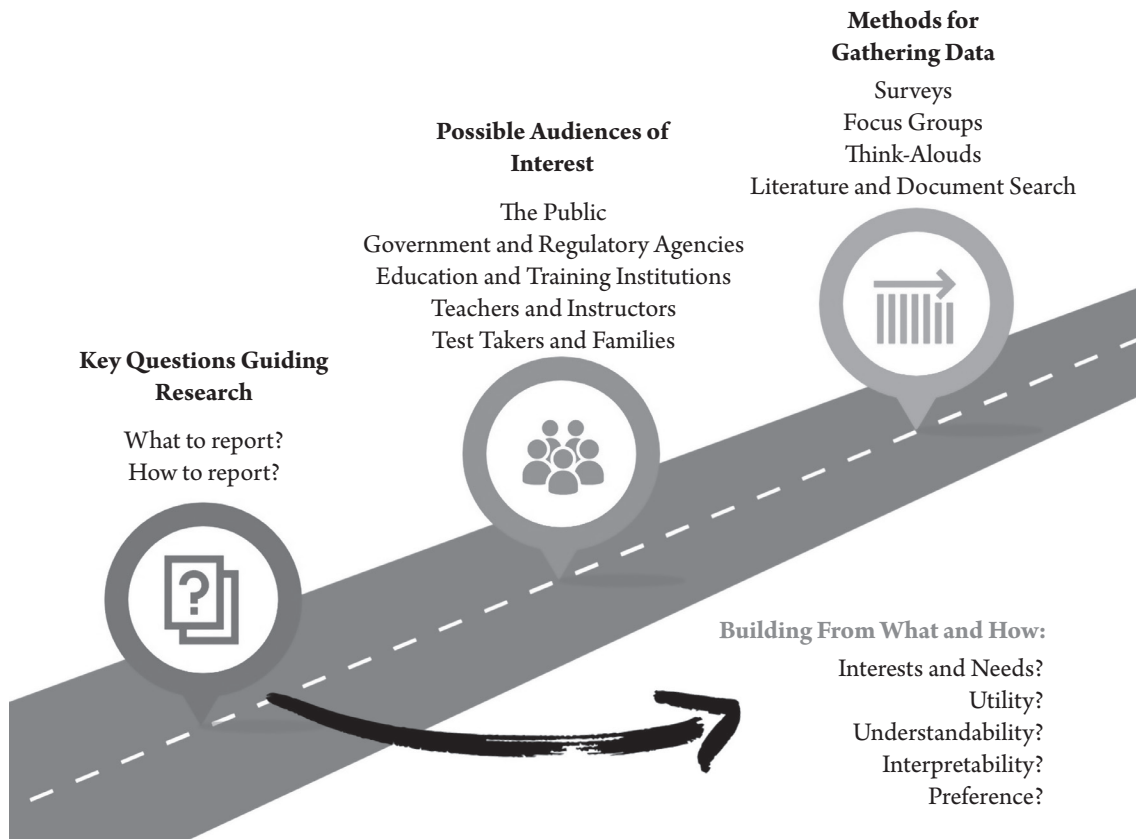


FIGURE 13.3
A Road Map for Audience-Based Research

Note. Adapted from *NAEP Validity Studies: Reporting the Results of the National Assessment of Educational Progress* (Working Paper 2003-11), by R. M. Jaeger, 2003, U.S. Department of Education, Institute of Education Sciences.

ensure that a key dimension of how report quality is defined is informed by direct communication with and solicitation of input from the intended users.

Another fundamental source for informing a working definition of an *effective* report is drawn from Hattie's approach working with educators, under the umbrella of his "Visible Learning" initiative (Hattie, 2009, 2010). In the realm of results reporting, much of Hattie's work has sought to elevate the importance of the audience and supporting stakeholders in using assessment data. The lessons of his research on learning, achievement, and results reporting transcend their original context to apply to all of the reporting contexts relevant here. In Hattie (2010), 15 principles for establishing the validity of results reports are enumerated, grouped in topics including validity of reports, sources of validity evidence for reports, and design principles for reports. Among these principles are the ideas that reports should have a specific theme and be designed to address specific questions posed by the stakeholder and that, from a conceptual point of view, reports should be conceived of as actions, not screens to print. These are critical principles that, similar to the ideas in Wainer's work, speak to the notion of what effective is that will guide this chapter.

Reporting and Professional Standards

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), offers an additional lens for evaluating report effectiveness for the contexts described in Figure 13.4: educational assessment, workplace testing and credentialing, and psychological assessment. The *Standards* are presented in clusters, and two clusters specifically address results reporting: one focusing on reporting and interpretation and another on test takers' rights to fair and accurate reports. The first standard from the first cluster is the most emblematic of the kind of recommendations provided:

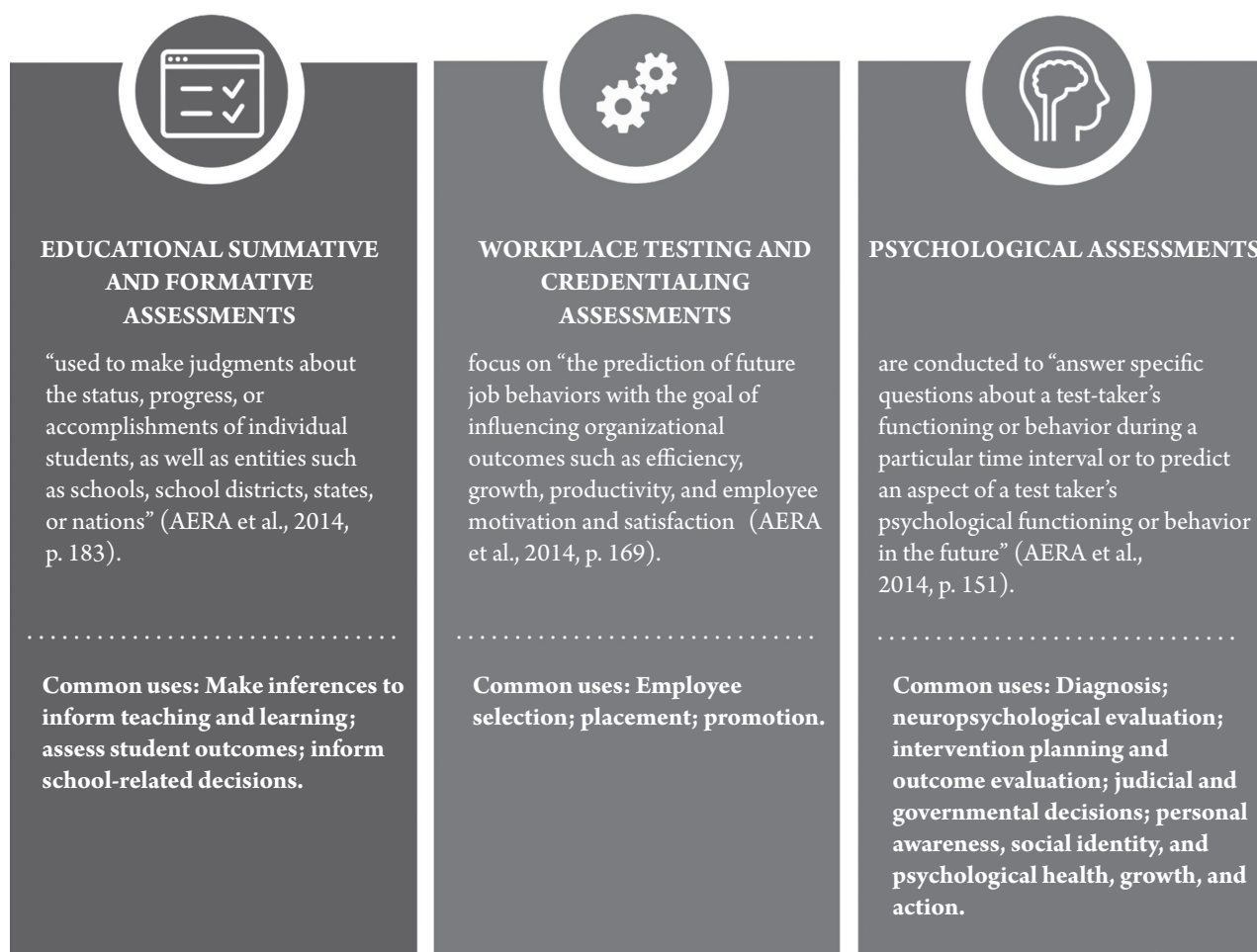


FIGURE 13.4

Overview of Assessment Contexts for Reporting, per the *Standards for Educational and Psychological Testing*

Note. Information is drawn from *Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014.

When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (AERA et al., 2014, p. 119)

With sentences such as “provide interpretations appropriate to the audience” in the text of the standard and “reports and feedback should be designed to . . . minimize potential negative consequences” in the accompanying commentary, the *Standards* offers reporting ideals in ways that are necessarily broad to account for guidance that spans highly varied testing purposes. Examples of how those ideals might be achieved are discussed, but the specific approaches in the examples are options rather than requirements, in alignment with the thinking that there is no one-size-fits-all report design that will succeed in every testing context.

Within the two clusters, there are standards addressing the need to ensure the validity of automatically generated interpretive text (Standard 6.11) and select performance labels that support intended inferences without being stigmatizing (Standard 8.7; e.g., O’Donnell & Sireci, 2021). There are numerous performance-level labels in use, as illustrated by the word cloud in Figure 13.5 based on O’Donnell’s (2020) review of labeling practices for statewide assessments in the United States (larger text represents more frequent use of a word). There are also standards for more procedural aspects of reporting, including report delivery and handling material errors. Beyond the standards in those clusters, O’Donnell and Zenisky (2020) identified 36 standards (AERA et al., 2014) that apply to results reporting, many of which are mentioned later in this chapter along with standards from the National Commission for Certifying Agencies (2014), which are similar but geared toward credentialing.



FIGURE 13.5
Performance-Label Word Cloud

The International Test Commission's Guidelines on Test Use (International Test Commission, 2013) are also instructive here: While some of those guidelines offer assistance in the selection of tests and test administration, Section 2.7 focuses on interpretation with 12 substandards pertaining to generalization of scores, reliability and validity of results and implications for meaning, and minimization of bias and social stereotyping that may negatively impact individuals and groups. Section 2.8 of these ITC guidelines further relate to communication of results, addressing privacy concerns and providing guidance about communication strategies that take audience interests into account. Overall, the guidelines from AERA et al. (2014) and other groups place validity and understanding one's audience front and center, promoting the idea that an effective report is one that guides users toward valid interpretation and uses, leaving it up to developers to find the exact content and design features that will be most successful in that task.

From Score Reporting to Results Reporting

Historically, the topic of communicating test results has been referred to as score reporting and, indeed, scores (primarily overall scale scores) were the main or only data contained on reports for many years. Over time, however, there has been a clear shift in most large-scale testing programs where what is being communicated to audiences is substantively different. Much of what is found in present-day reports extends well beyond simple scores to include achievement levels (e.g., *Pass, Not Yet Meeting Expectations, Proficient*), information to support comparisons to relevant groups (e.g., percentile ranks, reference group descriptive statistics), and numerical or categorical subdomain feedback. Other results sometimes included are item-level performance data, where item-level results can be provided for individual test takers or groups of test takers, and growth results (Zenisky et al., 2019), which use test scores from previous administrations to project test performance in the future.

Although it is certainly true that most of the data presented on current reports are scores or derivatives of scores, the shift in language being advanced here—from “score report” to “results report”—supports a more inclusive perspective. This change in emphasis is intentional in that it reinforces a more expansive approach to representing test performance. The task explicit in this shift is to consider and explicitly prioritize ways to communicate test information that are richer and more engaging. For example, there is growing research on interactive reporting tools such as dashboards, discussed later in this chapter.

Our recommendation, reflected in the terminology shift, is to approach the task of reporting assessment results as sharing multiple pieces of information with unique value rather than treating data other than the overall score as secondary in importance. This line of thinking may help develop reports that not only fulfill informational purposes and contribute more meaningfully to reporting systems, but also guide users to what comes next on the basis of the results, aligning with the third of Hattie and Timperley's (2007) reporting questions: Where to next?

PERSPECTIVES ON REPORTING FRAMES AND VALIDITY

Clear and useful score reports support users in making appropriate score inferences and have an important role to play as part of efforts to explain the validity argument for a test to key stakeholders. (Hambleton & Zenisky, 2013, p. 479)

Reporting as Validity in Action

Validity is the most essential consideration in the process of developing and evaluating tests. In turn, how test results are interpreted and used is central to how validity is defined in the AERA et al. (2014) *Standards* because validity there is understood as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). Interpretations and uses are also key to the argument-based approach to validation (Kane, 2006, 2015), which provides a practical framework aligned with the *Standards*. Following this approach, at the outset of test development, test developers must first create an interpretation and use argument (IUA), listing all proposed interpretations and uses along with the reasoning for each, and then proceed to a validity argument, which involves a strategic evaluation of the claims and assumptions in the IUA. The development of an IUA, then, can provide a clear evidentiary basis for reporting (Ferrara & Lai, 2016).

Similarly, based on the *Standards* (AERA et al., 2014), test developers are asked to articulate all intended interpretations of results for specified uses (Standards 1.0, 1.1, 11.1) and later gather and evaluate validity evidence for each of those interpretations (1.0, 1.2, 1.11–1.25, 11.1). Intended interpretations and uses are mentioned in standards regarding scores (5.1, 5.4), precision (2.0), norms (12.5), and various other topics. A smaller number of standards focus on ensuring that those intended interpretations come to fruition when reports reach their audience:

Standard 3.1: Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant sub-groups in the intended population. (p. 63)

Standard 13.5: Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied. (p. 211)

This notion of effective reporting is in line with a view of validity heavily informed by the work of Messick (1989), with an emphasis on consequences and use. Reporting is essential to validity because it is the means through which intended interpretations are realized, and this is critical to interpretation, particularly in terms of the intended (and unintended) consequences of testing (per Messick, 1989). The greater focus on *intended* rather than *actual* interpretations in the AERA et al. (2014) *Standards* is likely because the former are needed to start building an IUA and are available earlier in the test development process. However, the interpretations made by users on the basis

of reports should also be a major component of validity evidence. If a report fails to guide users toward the validated interpretations for an assessment, the chances that the assessment program will meet its goals are greatly diminished, regardless of how much evidence exists in support of ideal, but unrealized, interpretations. Ryan (2006) also espoused this connection between validity and interpretation, suggesting that elements or features of reports that foster or lead to unsupported interpretations have a corrupting effect on validity. (See also Lane & Marion, this volume.)

Tannenbaum (2019) described this issue in terms of alignment: “A score report that is not well aligned with the test is of little value; similarly, a score report that is well-aligned, but not communicated to users in a way understandable to them is of little value” (p. 9). This is a critical point that speaks to an important dimension of effective, in that a report without alignment or meaning is essentially devoid of consequence and occupies space without adding value. To ensure alignment, Hambleton and Zenisky (2013) proposed explicitly linking validity evidence and the purpose of a test to the contents of results reports and subsequently investigating the characteristics and information needs of the intended audience to determine how to best present that content (all steps of the Hambleton and Zenisky model are described in a later section). A common thread among Tannenbaum (2019) and Hambleton and Zenisky (2013) is the view that effective reporting is critical to validity, and ensuring that reports can be interpreted as intended to fulfill specific purposes is as important as having other sources of validity evidence.

Several authors have proposed alternate notions to better situate reporting as part of the validation process, recognizing its critical role. These include the following:

- The validity of reports: the notion that reports deserve their own validity arguments (Hattie, 2010);
- User validity: the perspective that the conceptualization of validity from the AERA et al. (2014) *Standards* should be expanded to include a user-centered source of evidence based on the accuracy and effectiveness of interpretations based on test output (MacIver et al., 2014);
- Report interpretability: the view that the interpretability of reports should be considered an aspect of validity (Van der Kleij et al., 2014); and
- IUA source materials: the view that reports and other technical documentation of tests are source material for the IUA for an assessment (Ferrara & Lai, 2016).

O’Leary et al. (2017) expanded on those ideas, proposing an approach that aligns with the unified conceptualization of validity (Messick, 1989). Instead of discussing new types of validity, O’Leary et al. (2017) recommended expanding the five sources of validity evidence—content, response processes, internal structure, relations to other variables, and consequences—to include interpretability: evidence focusing on the “adequacy, accuracy, and effectiveness of user understanding of scores and the consequences of testing” (p. 20). Further, they proposed that evidence based on interpretability and evidence based on consequences be treated as having equal importance as

the four other sources, which traditionally have only yielded evidence in support of possible interpretations.

Alternatively, the five current sources of validity evidence could be adapted to encompass results reports more explicitly. Tannenbaum (2019) provided several ideas of how this might happen. Evidence based on test content, which is usually about the relationship between the content of the test and its target construct, could also include evidence that the content of the report is well aligned with the content of the test. Evidence based on response processes, which focuses on how well the test elicits the strategies or cognitive processes that are key to the construct, could incorporate evidence that users visually navigate the report as intended, spending more time on the areas the design of the report was meant to highlight.

Evidence based on internal structure is typically about how well the relationship among items or portions of a test reflect its target construct and the extent to which those relationships stay consistent across subgroups of test takers. When applied to reporting, this source could involve gathering evidence of how well report users understand the ways in which different parts of the report relate to one another (e.g., how a sentence about precision should inform conclusions based on a graph with content area results). It would also be relevant to evaluate how different groups within the target audience interpret and use the information reported (e.g., groups from different socioeconomic backgrounds, test takers scoring below and above a passing score).

Gathering evidence based on relations to other variables in the context of evaluating results reports is more complex. Traditionally, this type of evidence comes from analyzing the relation between test scores and other measures. To evaluate reports, Tannenbaum (2019) suggested investigating the match between students' competency as communicated on their score report and as evaluated by teachers, noting that convergence would be confirmatory evidence and discrepancy would be more challenging to interpret (it could reflect differences in what was tested compared to what teachers considered in their evaluation).

Finally, validity evidence based on consequences of testing has a strong connection to reporting. The interpretations individuals make and any actions they take in response to results reports are a direct consequence of testing. So are potential misinterpretations, inaction, and undesired actions. This type of evidence, which could employ multiple methods, should shed light on the extent to which reports are interpreted and used as intended based on reasonable expectations or the IUA for a test.

Whether testing organizations plan test development and validity research in alignment with the argument-based approach to validity or a framework such as principled assessment design, described next, effective reporting should be treated as essential to validity and given due consideration.

Reporting and Principled Assessment Design

Principled assessment design (PAD) refers to several approaches, including evidence-centered design (Mislevy & Haertel, 2006) and assessment engineering

(Luecht, 2013), intended to offer a more integrated pathway for designing or improving assessments and building validity arguments. Huff et al. (this volume) offer a comprehensive overview of the framework, detailing how—in contrast to more conventional methods that may lead to siloed activities—it employs strategies to ensure coherence “from construct definition through task development and score inference” (p. 447).

Both PAD and the argument-based approach to validity reflected in the AERA et al. (2014) *Standards* rely on the explicit articulation of ideas to guide the design and validation process. However, while the argument-based approach calls for the articulation of proposed interpretations and uses of assessment results (along with their rationales), PAD requires integrated documentation and articulation of the assumptions and decisions related to *all* aspects of test design and validity research. This focus on approaching the assessment cycle in a more unified way is ideal for reporting work, which in the past seemed to be treated as an afterthought in the test design process (Katz, 2019).

PAD is a promising framework to help those involved in test development meet present-day challenges, including greater distrust and scrutiny of assessments and the increasingly common expectation that one assessment may serve multiple purposes (Huff et al., this volume). In practice, approaching results reporting in the context of PAD may take many forms. It may involve creating a prospective score report early on to promote discussions about what information the assessment should provide (Slater et al., 2019; Zapata-Rivera et al., 2012), using the prospective score report to clarify the claims to be made about test takers (Zieky, 2014), and changing test design specifications to align with the desire to report specific information, such as subscores (Sinharay et al., 2019).

What is common across these examples, and central to PAD, is the idea that reporting considerations are a key component of assessment design—not a step to be taken when other important test design decisions are already finalized. In such an unfortunate scenario, the report must conform to those established test design decisions, which may not be aligned with the information needs of report users. PAD offers an approach for considering test content, format, and reporting in tandem, making it more likely that results reports will be effective in promoting validated inferences. This idea of articulating reports and their purposes in advance might be viewed as backward design, in the sense of starting with the product and working through the process early on to get to that place in the end, but such an approach is quite forward thinking and in effect prioritizes understanding and actions associated with reports before it is too late to change course.

Reports as Data Stories

One of the ways of making statistical results more meaningful to intended audiences is to report the results by connecting them to numbers that may be better understood than test scores and test score scales. For example, to relieve the concern many persons had about flying after the TWA crash a few years ago,

the airlines reported that there is a single plane crash every 2,000,000 flights. In case the safety of air travel was still not clear, the airlines reported that a person could expect to fly every day for the next 700 years without an accident. Probably some people felt more confident after hearing these statistics reported in this way. Knowing that the probability of being in a plane crash is less than .0000005 may not be so meaningful. (Hambleton, 2002, p. 194)

Building on the approaches to validity and PAD referenced here, a key idea to raise at this point is that of the *data story*, where results reporting can be conceptualized as a coherent and planned approach to communication about assessment data, rooted not only in the validity of the inferences being made but also in the validity of the communication about the data. Data story is a term that has emerged from business and marketing settings in recent years to describe a highly coordinated strategy for talking about data in various nonassessment contexts, but, given the nature of data in reporting, there is a natural relevance to this topic in educational and psychological testing. The basic concept of a data story is that it is a thoughtful and intentional approach to sharing data that brings together data visualizations and compelling narratives, targeted to specific audiences, that aims to help intended users of the data understand the data and take action where appropriate (Hooper, 2021). It is not simply a matter of better (or different) graphics, but rather an orientation on the part of report developers to identify a story to be told with data and then use tools such as visualizations and one or more narrative patterns to explain something about the data (and why the data matters), in the context of a specific vehicle for communication (e.g., a results report, a presentation, or an online reporting tool).

Bach et al. (2018) defined a narrative—in this data story context—as giving shape to the *events* in a data story, following a specific narrative pattern. The pattern(s) that underlie any data story can vary from context to context, and the choices made are predicated on a narrative's intent. Narrative intent is an important idea here, in that it reinforces the purpose of the communication effort and can be linked to the context of assessment through articulation of both (a) test purpose and (b) report purpose. Examples of intents that might underlie data stories from the data story literature could include enlightening audiences, evoking emotional responses, spurring action, and questioning beliefs and behaviors. It is not hard to move from these general intentions to the aims of results reporting with respect to these intents.

To select and use any narrative pattern, the author must have a story idea, an idea of the intended audience, and an intended effect such as sympathy, action, information, or explanation (Bach et al., 2018). Conceptually, this links quite closely to the layers of reporting put forth by Behrens et al. (2013), in which three layers of reporting were articulated. Layer 1 corresponds to information communication (in effect, the “what” of reporting, asking, What is the story to be told?), while Layer 2 is couched in social activity (How are people approaching the story and who is the audience for the story?). The third layer of Behrens et al.'s (2013) approach aligns to societal transformation

(asking, What is the moral of the story?)—in effect, the intended outcome, or the *why* of the act of data communication.

Drawing on visualization and data story work, as well as arts and literature, there are several ways to think about narrative patterns. One overarching narrative strategy that has been identified in this area is explanatory versus exploratory orientations (Thudt et al., 2018). This has been discussed in the context of results reporting by Zenisky and Hambleton (2012b) relative to digital reporting efforts. Explanatory reporting efforts are predicated on the report developer's conscious and explicit choices about content, appearance, and interpretation support, while exploratory efforts typically align to interactive tools that offer users flexibility and personalization of the experience of the data story. Most traditional results reports in use for educational assessments in the early 21st century follow the explanatory narrative strategy, while publicly available anonymized databases on state websites draw on the exploratory model, with drop-downs and selection boxes that allow for customization.

Another narrative pattern to consider in the context of reporting involves prediction and how users might be presented with formative results to guide future action. Prediction, as a reporting aim, draws on the past to identify next steps. Thus, reports that address this aim may use data that might be presented in a static way using an explanatory strategy to set the stage for the results and include an exploratory component that offers the user the opportunity to navigate through to identify areas of strength and weakness and connect the results to specific actions.

At a similarly straightforward level, another approach to narrative patterns is to think of some narratives as linear (following a highly temporal or sequential route, such as the Harry Potter books tracing Harry's path from preadolescence to adulthood in the wizarding world) while others are nonlinear (such as the television show *Lost*, with the story told using flashforwards, flashbacks, and even sideways/parallel paths; or *The Godfather Part II* film, which juxtaposed events in the lives of the Corleone patriarch and his son to unfold both stories semisimultaneously). In the context of results reports, the analogue to a linear narrative pattern is the body of work that is produced to communicate the long-term trend results for NAEP (National Center for Educational Statistics, 2013). The overarching data story in long-term trend results is based on data that are linear in nature, with many visualizations used that aim to illustrate the performance of test takers over time.

Exploratory or explanatory and linear or nonlinear are not, however, the only narrative patterns that can be used for data stories. Bach et al. (2018) identified five different groupings of narrative patterns based on broad intent. These groupings are provided in Figure 13.6.

In reflecting on these groupings and the specific patterns within each one, it is important to note that any data story may well use more than one of these techniques to accomplish different intentions within the same document. For example, on a typical individual student results report for K–12 assessment in the United States, narrative patterns that can be spotted relatively quickly might include familiarization



Patterns for argumentation

These serve the intent of persuading and convincing the audience. Examples of these patterns are compare, concretize, and repetition.



Patterns for flow

These patterns serve the intent of structuring a sequence of a message or argument, setting the order, rhythm, and pace of the story. Examples here are reveal, slowing down, and speeding up.



Patterns for framing the narrative

Patterns here serve the intent of controlling how facts and data are perceived and understood. Examples of framing narrative patterns are familiar setting, make-a-guess, defamiliarization to challenge expectations/convention breaking, silent data, and physical metaphors.



Patterns for empathy and emotion

These serve to engage the audience with the content of the data story. There are overlapping narrative patterns here (reveal, slowing down, speedup, and concretize). There are also some novel patterns as well, such as breaking the fourth wall, humans behind the dots, and familiarize.



Patterns for engagement

This category serves the intent of connecting the audience with a story, to make them feel part of it and perhaps even offer them some control. Specific tracks that can be taken here include the use of rhetorical questions, call-to-action, make-a-guess, and exploration.

FIGURE 13.6

Narrative Patterns for Reporting Data Stories

Note. Adapted from “Narrative Design Patterns for Data-Driven Storytelling,” by B. Bach, D. Stefaner, J. Boy, S. Drucker, L. Bartram, J. Wood, P. Ciuccarelli, Y. Engelhardt, U. Köppen, and B. Tversky, 2018, in N. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale (Eds.), *Data-Driven Storytelling* (pp. 107–133). CRC Press. <https://doi.org/10.1201/9781315281575-5>

(using the student’s name in personalizing the report), compare (to compare the student’s observed score to a class, school, district, and state), and physical metaphors (with stoplight displays or pie charts to illustrate mastery of subdomain skills). This is quite common: Bach et al. (2018) commented on a number of use cases where multiple narrative patterns are used in data stories, pulling examples from different media sources.

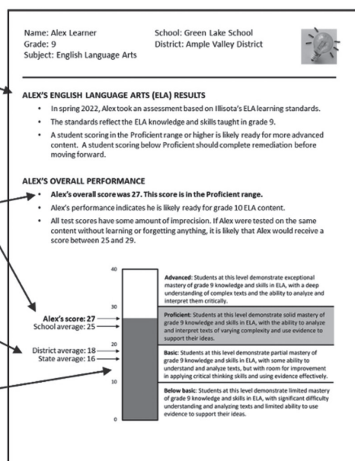
The literature on results reporting itself offers a number of additional narrative patterns. One such approach is *question and answer*, where the flow of the report document is sectioned by a series of questions formulated by the report authors to reflect common questions from intended users, with answers provided. Another approach is reflected in the traditional individual results report that presents a high-level overall result, followed by a progressive “drill-down” to more granular series of data points. In this case, the pattern of the data story moves in a linear fashion on the basis of the granularity of the data (from least to most), typically drawing in some comparisons and making some inferences highly concrete. The report might then conclude with a call to action (in a section entitled “Next Steps”). Figure 13.7 highlights the narrative patterns used in the sample report presented at the beginning of the chapter.

Empathy: Familiarize
Introducing the test and learning standards in the context of Alex's education, making them more relevant and relatable

Flow: Repetition
Using repetition to emphasize Alex's overall score

Argumentation: Compare
Showing Alex's performance in relation to three groups, reinforcing the overall result

Argumentation: Concretize
Showing both the score scale and Alex's score graphically, making them more meaningful



Flow: Reveal
Throughout the report, presenting information gradually (from broader to more specific) for better flow

Framing the Narrative: Convention Breaking
Adding a "warning" symbol to the Below Mastery result, drawing attention to it

Engagement: Call to action
Explicitly referring to "next steps" to promote action

Engagement: Rhetorical Questions
Posting and addressing questions that report users might have about the results

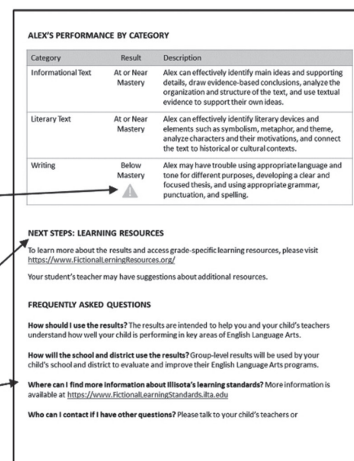


FIGURE 13.7
Simplified Sample Report With Annotations Highlighting Narrative Patterns

Hattie's (2010) work on the validity of reports likewise offered some guidance that aligns well with the idea of results reports as data stories. Principle 3 ("Readers of reports need a guarantee of safe passage") and Principle 4 ("Readers of reports need a guarantee of destination recovery") align to the underlying concepts of Bach et al.'s (2018) patterns for flow. Principle 5 ("Maximize interpretations and minimize the use of numbers") is a way into patterns for argumentation, and critically, Principle 8 ("Each report needs to have a theme") brings us full circle on the very idea of data stories and intent: defining the aim of the report and articulating the structural and communication techniques that can be employed to achieve the intent of a specific data story, for specific user groups.

The value of the data story and narrative pattern framework lies in the premise that reporting in educational and psychological testing is an intentional act of communication about data, and the approaches to conceptualizing and organizing the stories discussed here, drawn from outside measurement, have significant interdisciplinary value. Across assessment contexts and reporting settings, there are many data stories to be told (at both individual and group reporting levels) and many ways to accomplish the communication of assessment data. Under this paradigm, report developers must look at the data story of the report as a coherent whole, as well as how distinct report elements contribute to the story in a purposeful and organized way.

ESTABLISHING REPORTS, REPORTING SYSTEMS, AND REPORTING CONTEXTS

There is no one-size-fits-all design for what constitutes a "good" score report. Tests and test purposes are different, and users' data needs vary. No one magical visual display will make test results understandable to all users, and—sadly—there

is no one perfect line of text that illuminates what standard errors are and why they matter. (Zenisky & Hambleton, 2015, p. 586)

In considering the breadth and depth of what reporting is more fully, it may be helpful to extend the metaphor of the data story to assist in defining critical reporting dimensions and aspects. Specifically, the questions of *who*, *what*, *where*, *when*, *why*, and *how*, long associated with journalism, can serve as something of an organizing framework for conceptualizing and disseminating results reports broadly in the context of assessment, in line with the data story metaphor.

Who?

The first of these organizing questions is *who*. In results reporting, there are in effect two very significant ways to address this fundamental point in report development. One of these involves the subject(s) of the report, in terms of whose results are being reported. Quite simply, a report can be crafted to describe the performance of one (a single individual) or many (more than one). From the perspective of developing a report, this first aspect of *who* concerns the unit of analysis for the report—individuals or groups (and how group membership is operationalized for the purpose of the report).

From that natural division arises the second way to address the *who* of reporting, which is the intended user. Reports are developed to describe the performance of an individual or a group on a specific assessment, and building on that, individual and group reports are typically developed for different intended audiences (to fulfill different reporting aims—the *why* of reporting).

The most detailed reports for individual test takers are typically disseminated to the test takers themselves and, depending on the testing context, others close to them in the social space, such as family members and educators. Indeed, reports for individuals are typically more relevant for test takers and their own families and for the test takers' instructors. In the context of credentialing, individual reports are provided to the test takers themselves, and the considerations of reporting are differently nuanced between the personal and the professional uses (O'Donnell & Sireci, 2019).

As the social distance from the test taker grows, the nature of the substantial detail around the performance of the individual can change. In the example of certification and licensure, individual results at the global level of pass or fail may be transmitted to a local or national governing body for record keeping and maintenance of professional standards, but the focus in this group-level report to this user group is not instructional or formative, and as such, subdomain results or other score breakdowns may not be of primary interest for these particular users. (Detailed results exploring comparisons between test-taker subgroups will be of interest, but such data are typically summarized and anonymized for governing body uses [record keeping, monitoring, and public communication].)

To that end, results for groups of test takers can be constructed to reflect groups of hyperlocal interest (such as a class, grade level, or school) all the way to aggregations

for national and/or international comparisons. School and district personnel may use group-level reports to identify academic strengths and weaknesses within classes, schools, and districts, and at the larger scale (such as states) statistical power can be used to help identify broader trends in student performance over time (Zenisky & Hambleton, 2015). Reporting units of interest can be based on geography (such as class, school, district, state) or other relevant demographic variables (e.g., race/ethnicity, language status, individualized educational plan status). Thus, results reports for groups can be reported on the basis of groupings constructed to reflect geography, demographics, and/or other dimensions of interest. The users of various group reports can be close (as in educators and coaches) or more removed (such as policy makers, the general public, and the media). The more removed such users are, the larger the aggregation (typically), and the purpose becomes more policy oriented in nature.

Though sometimes it is the individual score report that garners much of the interest in terms of reporting research, a great deal of foundational work on reports and reporting quality occurred within the context of evaluating NAEP group-level results (Zenisky et al., 2016), where research focused on interpretation of the reporting scale (Beaton & Allen, 1992; Hambleton, 1998, 2002; Hambleton & Meara, 2000). Other work prioritized audience-specific differentiation (Forte Fast & Tucker, 2001; Jaeger, 2003; Levine et al., 1998; Simmons & Mwalimu, 2000), displays for state results and other subgroups (Hambleton & Slater, 1997), market-basket reporting (Mislevy, 1998; National Research Council, 2001), and digital reporting strategies (Zenisky & Hambleton, 2007; Zenisky et al., 2009). All of these efforts have contributed in significant ways to the present understanding of report development processes and report evaluation procedures.

What?

Test scores are elusive. Even the popular percent score scale that many persons think they understand cannot be understood unless (1) the domain of content to which the percent scores are referenced is clear and (2) the method used for selecting assessment items is known. (Hambleton, 2002, p. 194)

The *what* of reporting, naturally, is about the contents of reports. Fundamentally, this is a question about data and what data elements are being communicated. It is a broad question, not only incorporating the specific results reported but also encompassing the presentation and interpretation supports (text and visual design) that play a critical role in shaping how results are communicated via reports.

As an activity that occurs part and parcel in the progression of test development, reporting encompasses the process by which data about human knowledge and skill (as measured by an instrument of some kind) are conveyed to intended users (Zenisky, 2015). In this way, the answer to the *what* of reporting is not simply a matter of scale scores, performance levels, and average scores for reference groups, but rather can perhaps be better thought of as a reflection on the key components of a report, in the sequence of the data story, referenced earlier.

In any data story, context is important, and a first way to address *what* in reporting involves the larger backdrop of the assessment context and what information broadly matters given the assessment being reported. For the purpose of results reporting, there are several key settings for reporting that are directly relevant, as previously presented in Figure 13.4: educational testing (summative and formative), workplace testing and credentialing, and psychological assessment. Each of these types of assessments accomplishes different measurement goals, and accordingly, the reports for each vary significantly in terms of their content because the underlying data and use for each differs from one another, in important and meaningful ways.

Reporting is not just about scores and data, but rather begins conceptually with information and communication. Whether the aim is content knowledge at an endpoint or at places along an educational path, demonstration of mastery to enter a profession, or an indicator of status relative to a psychological construct, these conceptualizations help to focus the reporting activity and ultimately the report product that is composed of scores and other results.

Some of these types of reports serve primarily to report an informational purpose, while others are developed with the intent to guide specific actions, and importantly, this often connects back to the purpose of an assessment. For many summative assessments in the educational context, the main reporting aim is status, and that is therefore reflected in the statement of purpose, though there may be a secondary nod to next steps in the statement of report purpose and in the report itself. In certification and licensure, the assessment purpose is to determine qualification to hold a specific credential, so reporting there prioritizes that information, though an exception here can be for failing candidates. At the discretion of a testing agency, candidates who do not pass may receive some guidance about relative strengths and weaknesses, from which the candidates can infer what might be valuable to focus on during future test preparation efforts. Formative assessment, as well as diagnostic assessment, both function at something of a crossroads in this regard. The typical aim of formative and diagnostic assessments is to inform instruction and identify areas of need, so while the reports contain information and the user of the report may or may not act on that information, the intent of the report is to provide actionable next steps. In this respect, the extent to which the aim of a report is fulfilled is up to the user.

To that end, briefly, a typical report begins with introductory content of some kind, to set the stage for the data being communicated. The nature of this introductory content may vary considerably, from individual reports to group reports, and depending on the audience. It may be as simple as a title that is descriptive of the results and some method of documenting whose results are contained within the report (as in the name of a specific individual or the parameters for inclusion in a group report, relative to geography, demographics, or other considerations). Some individual reports will go further in this regard, however, with text that describes the purpose of the assessment as well as the purpose of the report.

Another dimension of introductory content may include a statement to the effect of what is assessed, aligned to the purpose of the assessment. This should offer some broad contextual information as to the nature of test content and what is assessed. As reported by Zenisky and Hambleton (2015), a noninclusive list of examples of the descriptive elements of reports may include information such as:

- test name and/or test logo
- test date
- report title
- report purpose
- test purpose
- introductory statement from testing agency or governing body personnel
- individual reports: header space with identifying details such as name, address/school, group membership or status (individualized education plan, language, etc.)
- group reports: header space with identifying details for reported group(s) including demographic, geographical, and/or other grouping variable specifications
- details for external links to additional resources, such as curriculum materials and interpretive guides
- information about the location of frequently asked questions documents or other resources for score inquiries
- guidance on test score use/links to interpretive guidance
- glossaries of terms

The next section of most score reports is the high-level, overall, or primary results. On an individual results report, most agencies provide the overall scale score and—where appropriate—the performance level associated with that score. The formatting of these results can vary considerably, where different agencies may use different visual strategies to communicate the overall results data for one or more content areas, such as typography and font size, but also tables, bar graphs, and/or line graphs. For group reports, the overall data may be presented in a summary form, using a table or a graphic structure to communicate specific points about the data. The high-level results, whether on an individual or a group report, in the general case typically provide the results that align best to the primary purpose of the assessment. Below is a nonexhaustive list of types of scores that can serve as the basis of the scores reported (Zenisky & Hambleton, 2015). Some of these types of scores are more common in some testing applications than others:

- raw scores
- scale scores
- percentage correct
- percentiles
- stanines

- grade equivalent
- *T* scores
- performance classifications (e.g., *Advanced*, *Proficient*, *Basic*, and *Below Basic*)
- subtest or subdomain scores
- item-level scores
- student score growth

As part of the overall results, many individual reports will include additional contextual data in the form of comparisons to relevant reference groups. For educational tests, this typically means that the results for the individual student of interest may be presented alongside average scale score results for similar test takers in relevant geographic units (class, grade, city, state, country) or demographic groupings. Though many tests in the early 21st century are criterion-referenced and the primary interpretation advanced by agencies is performance relative to standards or benchmarks, these pseudo-norm-referenced comparisons are common, and often even expected. Many report users desire information about how the subject of an individual report is doing relative to others, even if little or no data about the distribution of performance are provided. It should be noted that in recent years, educational assessments have moved away from norm-referenced to criterion-referenced tests, but some form of normative data as described here is still present on many results reports. A number of examples of individual student reports can be found on the current websites for U.S. state education departments, as well as within Goodman and Hambleton's (2004) paper.

The next section(s) of many reports offer finer grained results. It is in this section of reports that, when these data are provided, report users begin to understand their performance on a test and why they achieved the score or result that they did. These finer grained results are what users look to in order to improve. The typical format of these results is the presentation of subscores, but this is a challenging topic for psychometrics because of the issues of dimensionality and the reliability of subscores and similar subdomain-level results. Most summative assessments are simply not built to offer high levels of reliability at the subdomain level, because they are intended to offer reliable results at the overall level given their articulated testing purpose and do not contain sufficient items to attain psychometric reliability for subsets of items. This is an active area of psychometric research (see Sinharay et al., 2019, for a thorough review of the considerations and issues), but a few recommendations from Sinharay et al. (2019) include the following:

- When possible, approaches such as evidence-centered design or assessment engineering are preferable where subscores are to be reported to address issues around the relative distinctiveness of the subscores since they conceptually support differentiation, because content blueprints alone may not avoid high correlations among subscores (Sinharay et al., 2007);
- Reported subscores should exhibit psychometric properties of adequate, reliability, validity, and distinctiveness;

- Agencies may wish to consider using weighted averages or augmented subscores for diagnostic purposes, rather than subscores to address the quality considerations of reliability, validity, and distinctiveness;
- When sufficient items are present for subscore reporting, subscores can be reported on the scale score metric for a test, and subscores can be equated so that interpretation of subscores can be interpreted and used consistently over forms;
- The quality of subscores aggregate over test takers can be useful but should be evaluated for indicators of quality (reliability, validity, and distinctiveness); and,
- The expectations around subscore quality for large-scale summative assessments can and should be applied for formative assessments.

It should be noted that many reports do include some form of subscore or subdomain results to provide a minimum of guidance for test takers, particularly those who obtained lower scores and who may seek information on how to improve. An additional set of recommendations around making choices as to which subscores to use and when from Sinharay et al. (2019) are as follows:

- Reporting of observed subscores may be reasonable if the subscore is reliable and dimensionally distinct.
- If evidence for subscores does not support their reliability and distinctiveness, then no amount of statistical adjustment can address fundamental limitations of the scores, and thus reporting of subscores cannot be justified.
- If subscores are moderately reliable and moderately correlated, then statistical adjustment may help.

Whereas simple subscores are easy to compute and are easily understood, Sinharay et al. (2019) do provide a comprehensive list of techniques to compute and adjust subscores, as referenced above, which can be drawn on to address the unreliability and lack of distinctiveness problems. These include augmented subscores and weighted averages, Yen's Objective Performance Index (Yen, 1987), the use of cognitive diagnostic models, multidimensional item response theory (IRT) models, and scale anchoring (Beaton & Allen, 1992). It is also certain that as data mining and data analytics work advances, so too may new methods for subscore reporting emerge.

Clearly, this topic, as an issue of psychometrics and operational practice, is not a settled matter. Large-scale assessments rarely, if ever, meet the numbers of items needed to obtain high values of traditional reliability in subscore reporting. For example, Sinharay et al. (2019) suggested that a threshold of 0.8 for reliability (based on Nunnally, 1978) is reasonable, and to get there might require 20 items. This raises philosophical issues of (a) what constitutes quality in subscores (when viewed through the lenses of reliability, validity, and distinctiveness and how those are defined) and (b) potential for harm (what are the consequences of reporting subscores that do not meet such thresholds/definitions of quality?).

Detouring briefly to group-level reports, formats typically include list-style reports of individual performance or reports that highlight summative statistics that describe

performance in aggregate. It is important to note that as a matter of good practices, the *Standards* (AERA et al., 2014) do suggest that contextual information is critical for reporting on groups, especially when reporting on group differences. For example, Standard 12.17 suggests that reports should include relevant contextual information to support meaningful and appropriate interpretations. Noted here as well is a hybrid approach to group reporting when such results are accessed online. Often such reports are targeted to instructional users and mix levels of aggregation. For example, a class-level report for a teacher may have class-level aggregated results but also display a roster-type view. In that roster view, a teacher might want to carry out certain functions, such as sorting a student list according to performance on the assessment, or click on individual student names and see more detailed individual results. The final section of a report often includes some kind of concluding information and/or guidance for next steps. Once a recipient of a report has that report in hand (physically or viewing onscreen), many reports conclude with a brief section on what might be next for the user, particularly in the context of the individual reporting. Such report sections might offer some kind of summary about performance and links to various types of resources available locally or online. The idea is that once the report document has provided some information about the performance of a test taker, the stakeholder may, for example, want to find out more about the test and use the overall and subdomain results to collaborate and develop a plan for next steps and potential improvement. Similarly, this concluding section of reports might also have links to other external documents, such as frequently asked questions where stakeholders can access additional information. Whatever such documentation is linked to from a report, whether it be lesson plans, other instructional materials, or guidance of an administrative or informational nature, it should be noted that it is incumbent on the report agency that is promoting those materials—by means of such links—to articulate the purpose of such materials; establish standards for their quality, accessibility, and usability; and provide users with logical connections to align results with provided materials, where appropriate.

Concentric Circles of Results

One visual way to consider the *who* and *what* of reports, spanning the range of individual and group reporting efforts previously discussed, is shown in Figure 13.8. This idea of the concentric circles of interest in assessment results emerges from the audience-specific considerations and research road map illustrated in Figure 13.3, and as such, Figure 13.8 further provides a general framework for thinking about the progression of audience and results. It is important to understand that not all circles of interest are necessarily applicable for all tests. Beginning at the center of Figure 13.8, the results reported are at the most local and personal level—the individual test taker. A report at that level is crafted with the test taker as the reporting unit of interest, developed for the test taker (and perhaps their family, depending on testing context), and the report contents are scores and other forms of results, often linked to targeted or specific next steps of some kind. Moving outward, the next circle of results is intended at the level of

Who?

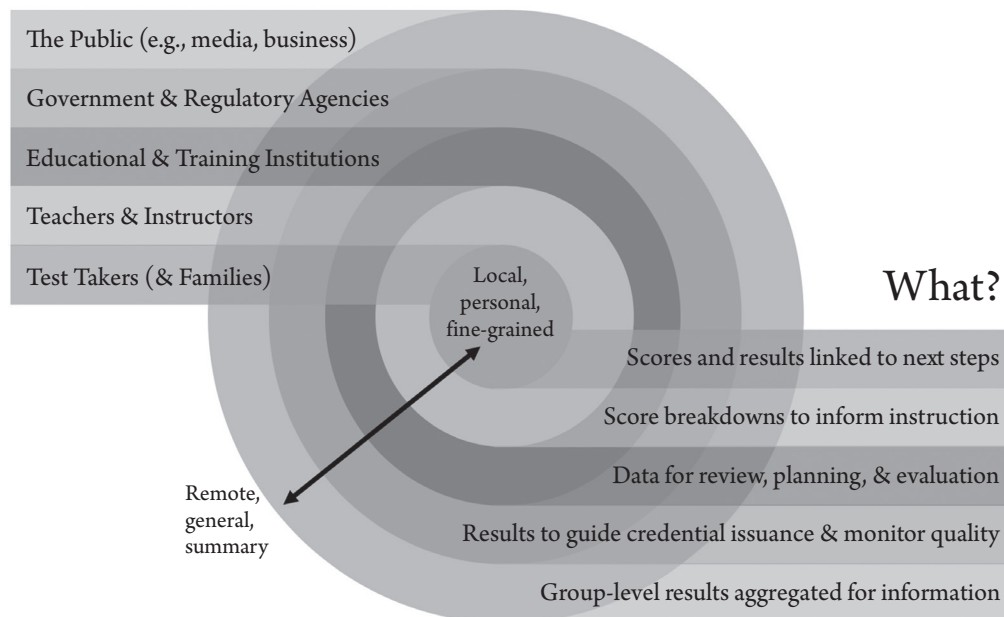


FIGURE 13.8
Concentric Circles of Interest in Assessment Results

teachers and instructors. Results at this level are still fairly local and focused on the individual, but also start to include relatively small but meaningful groupings (meaningful as defined by the user) given uses appropriate to that teacher/instructor role.

The next level of the concentric circles in Figure 13.8 is that of a progressively more general level of results, where results are communicated at the level of educational and training institutions. Here, the specific nature of individual results, and breakdowns of individual results, are typically far less consequential than a view of the results as data in aggregate to carry out goals such as review, planning, and evaluation. At this level, the performance of groups of test takers is typically analyzed and actions are identified based on the results.

Beyond that, moving further outward is the governmental and regulatory space for reporting, where individuals primarily matter for the (necessary and important) purpose of issuance of credentials. More of the results of interest at this level are for groups, with the assessment data fulfilling an informational/status purpose. Very little at this level is fine grained—the focus is on overall performance classifications for individuals and groups. The final, outer circle of interest is the public-facing side of assessment. This is the level of the kinds of databases and analysis tools such as the Data Explorers of TIMSS, PISA, PIRLS, and NAEP. Assessment results are completely scrubbed of personal identifiers and reporting is only at an aggregate level, with sample size suppression rules or bootstrap sampling methods used to shield the identification of individuals.

The overarching idea of these concentric circles is to encourage researchers and report developers to think about reporting efforts as a progression in communication. Reporting, for many tests, begins with the test taker and extends outward to other stakeholders. The nature of the results of interest necessarily change as stakeholders change, because the *use* of the results is different at each level.

Where?

When applied to results reporting, the question of *where* can be applied to format, as in, “Where are the results?” For many test takers, their results are detailed in the individual report, which is crafted to provide individual-level results in a highly structured way. Typically, such reports are static in the sense that whether disseminated on paper or onscreen via email or a secure portal, they are delivered as tightly arranged documents, with a specific sequence of results and information for every test taker, following the issuing agency’s template for the data story of interest. Rarely are individual reports for large-scale summative assessments marked by any measure of user choice in content or appearance; that is increasingly common for formative assessments.

This stands in contrast to group-level results: Though historically group reporting has filled volumes of expansive reports (e.g., the “Publications and Products” library of NAEP, <https://nces.ed.gov/pubsearch/getpubcats.asp?sid=031>), in recent years online tools for group reporting have proliferated (e.g., the NAEP Data Explorer, <https://www.nationsreportcard.gov/ndecore/>). In the former case, prospective users of the information obtain PDFs of static, large-scale results written by issuing agencies for reading and reviewing, while in the latter case, the users are in the position of exploring the data within the parameters of the publicly available tools, to answer their own questions about the data by, for example, selecting who is included, what results are of interest, and how those results are displayed (Zenisky & Hambleton, 2012b).

Interactive group-level reporting tools such as the NAEP Data Explorer and the International Data Explorer (<http://nces.ed.gov/surveys/international/ide/>) offer unprecedented customization, placing users in charge of pursuing their own data stories. This shift in how the reported narrative is constructed can be useful and empowering, especially to data-savvy users. That said, it can also be overwhelming and pose challenges to promoting appropriate interpretations of results. Those involved in developing online reporting tools should use tutorials and design elements to guide users and restrict access to unreasonable variable selections or analyses. Other strategies in this regard might include establishing a process for receiving and responding to inquiries, displaying data in multiple formats to support different processing preferences, and using a streamlined interface to the extent possible (Zenisky & Hambleton, 2007, 2012b).

One type of reporting tool that has been gaining traction is dashboards: highly visual interfaces that provide at-a-glance views of key information, often with some interactive features. Most dashboards currently available summarize learning analytics data for teachers (i.e., metrics of student progress and engagement, like time spent on online activities, grades, and number of outputs, such as discussion posts), but a growing

number are designed specifically for students (Schwendimann et al., 2017). The use of dashboards for reporting is variable: While some groups have been using them for many years (Brown, 2001), others have been more recently deployed in educational settings, as data mining, data visualization, and web-based tools have emerged in various disciplines to fulfill visual and functional aims in terms of data communication (Sarikaya et al., 2019).

There is growing research on ways to leverage learning analytics and the interactivity of dashboards to provide more customized results and recommendations (Verbert et al., 2013). Within the context of a formative assessment integrated with instructions such as Brightpath (<https://www.brightpath.com.au/formative-writing-assessments/>), users can interact with dashboards that connect information to specific next steps with real-time updating of student performance and even teacher evaluations. Another interesting application for dashboards is the use of open learner models, which are visual representations of what a student knows, as well as areas of difficulty that can be updated dynamically as more data become available (Bodily et al., 2018; Kay, 1997; Zapata-Rivera, 2021). For instance, a student receiving a static Grade 7 mathematics report may see that “understanding statistical variability” is an area for improvement and make a mental note of this result. In contrast, a student accessing a dashboard with an open learner model showing the same result could interact with the model by opting to answer additional questions to support their learning and, in turn, update their status on that content area (this interactivity is what makes the model open).

As a specific format within the realm of results reporting, with additional research dashboards can become a popular alternative or complement to static reports as consumer-facing, web-based applications with powerful data aggregation features that offer real-time indicators. Consider, for instance, the dashboards from exercise- and sleep-tracking apps accessed daily by millions of smartphone users, to obtain at-a-glance reporting of quantitative measures of health up to the minute of checking (e.g., steps taken, hours slept, heart rate). A key feature of dashboards, then, is this idea of real-time reporting, and in that respect data dashboards are perhaps best suited for use in results reporting contexts aligned with formative or diagnostic initiatives that update regularly and relatively frequently.

Since dashboards for assessment results offer a middle ground between static reports and highly customizable reporting tools and serve to meet a specific real-time reporting goal, it is important to determine which elements should be interactive and which should not, and this should align with the goals of the assessment program. Developers should also consider general guidelines from the dashboard literature, such as using customization to promote a sense of empowerment and agency (e.g., allowing students to set their own performance targets and to decide what normative information they want to see, if any), designing each element to inform a particular decision or set of decisions (e.g., ensuring each dashboard section has a purpose, such as deciding which content areas warrant additional review), and using multiple visualization approaches

(e.g., color, spatial position, and motion) to support rapid perception (Bennett & Foley, 2019; Few, 2006; Park & Jo, 2019). Last, those involved in creating dashboards may adapt recommendations and approaches from the results reporting literature, combining insights from broader dashboard research with specific guidelines for reporting assessment results (Corrin et al., 2019; Kannan & Zapata-Rivera, 2022; O'Donnell et al., 2021).

When?

In the terms of the data story, the *when* of results reporting is a fairly straightforward matter of now or later. Now, in this sense, is immediately at the conclusion of test administration. With the widespread use of computerized administration, testing agencies can provide test takers with their results before they leave their seat, displayed onscreen, and/or with a report of scores emailed to them, provided that the necessary scoring procedures and quality control checks are done in those brief moments. The “later” of score reporting can range from a few days to a few months because agencies may need time to score and verify performance. In large-scale summative educational testing in the United States, for example, it is not uncommon for results to be made available to districts after 2 to 3 months and perhaps take an additional 2 months to be distributed to the students themselves.

The dimension of *when* in reporting varies considerably depending on testing context. In credentialing, the results for a clear majority of tests are reported relatively quickly, where candidates receive their scores immediately or very shortly after the test session, and results are likewise seamlessly transmitted to certification and licensure bodies. In education, formative tests tend to report scores quickly, but as described, summative test reporting tends to take a great deal longer. The time-consuming nature of scoring (and, hence, delays in reporting) can be attributed to a few causes, such as the need to human score certain item types, to perform psychometric analyses, and to do other quality checks on the items and the scores, but such delays in reporting results are also a relative impediment to the use of scores (Brown et al., 2019). Per the *Standards* (Standard 8.8), the professional guidance around the *when* of testing indicates that results should be provided in a “timely” manner, but what constitutes timely may be a matter of debate and can be viewed in light of official test purposes. In some testing contexts, reports of results are expected and used quickly (such as entry into a profession), and in other contexts, longer intervals between testing and reporting may be more tolerated.

Why?

Arguably, the most interesting question of the data story metaphor as applied to results reporting is *why*. In the *Standards* (AERA et al., 2014), Standard 5.1 speaks to the fundamental requirement that test users should be provided with clear expectations of the characteristics, meaning, and intended interpretations of scores, but it falls to two other standards (8.8 and 9.16) to establish the underlying expectation that scores themselves are reported, particularly with respect to test takers.

Standard 8.8: When test scores are used to make decisions about a test taker or to make recommendations to a test taker or a third party, the test taker should have timely access to a copy of any report of test scores and test interpretation, unless that right has been waived explicitly in the test taker's informed consent document or implicitly through the application procedure in education, credentialing, or employment testing or is prohibited by law or court order. (p. 136)

Standard 9.16: Unless circumstances clearly require that test results be withheld, a test user is obligated to provide a timely report of the results to a test taker and others entitled to receive this information. (p. 146)

These two standards set a baseline for reporting to test takers, conditional on use (per Standard 8.8) and circumstances (Standard 9.16). If scores have implications for an individual, then individuals must be provided with their results report in a timely manner. Taking Standards 8.8 and 9.16 along with Standard 5.1 (described previously), it is clear that when scores matter, they must be provided along with guidance around interpretation and use.

However, the process of reporting results happens in line with test purposes (e.g., educational summative, educational formative, credentialing, admissions, psychological). This is an important point to maintain in the conceptualization and development of results reports for various intended audiences. Often, the *use* of scores falls to entities other than the test taker, and accordingly the nature of the report sent to those external entities hinges on what they plan to do with the data. In many—not all, but many—settings, reporting to the actual test takers simply involves a report that is informational in nature, and use is left to others. To this end, in those cases, the *why* of reporting depends on the intended user (the *who*). In effect, reporting to test takers in many cases fulfills the question of *why* with an informational aim, rather than an actionable use by the test taker. Then, for other, external audiences, the *why* of reporting is generally more aligned to the fundamental purpose of the test (to demonstrate competence in a profession, to rank order individuals to inform admissions decisions, to determine mastery of content for granting a diploma, etc.).

The main exception to this, of course, is in education. The reality of reporting in the 21st century is that most users of test data in educational settings have expectations for reporting that extend deeper into the data, beyond high-level performance characterizations and toward formative or instructional purposes, which are not purely informational in nature, nor are they often the validated purpose(s) of the test (see Brookhart & DePascale, this volume)—they are *actionable*. The *why* of reporting to these users is different yet again, in that their needs and interests in the data are centered on instructional planning and support at micro (individual) or macro (group) levels (Brown et al., 2023).

How?

Discussion of the methods and practices of score reporting would not likely be complete without some attention paid to the *how* of reporting, in this case meaning the

psychometric methods that are used to produce the scores and other results provided in reporting documents. Such methods can span from relatively simple descriptive values to scores that are obtained through complex statistical modeling approaches. In some ways, this is closely entwined with the *what* of reporting, in that it speaks to how scores and other results may be computed from test data. Of course, the *how* of obtaining scores depends on the source data and the reporting aim for those particular data. A few strategies for producing scores and other results are provided in the following pages, relative to reporting aims, reflecting a broad continuum of computational and conceptual strategies.

Raw scores are perhaps the simplest of approaches: They represent the number of points obtained out of the number of points possible, with no weighting or transformation done. Raw scores are often used in informal or classroom settings, but are not typically used to report overall test results on large-scale standardized tests for various reasons, including to ensure comparability of scores across forms where minor difficulty differences are detected or to minimize preconceived notions of the reporting scale (Hambleton & Zenisky, 2003). Some large-scale assessments have reported subdomain results using raw scores—or a simple derivative thereof, such as percentage correct—per Goodman and Hambleton (2004). Scale scores are a step up from raw scores in that some type of transformation is applied for reporting purposes, such as linear or nonlinear. Raw scores, and scale scores, can easily be averaged over test takers for local or more global indicators of performance, and in this way scores for individuals can be interpreted in norm-referenced ways against sample or population averages. Similarly, cut scores can be computed and applied to raw or transformed scores for grouping test takers into achievement levels (such as pass/fail or gradations of proficiency, such as those used in educational assessments and represented earlier in Figure 13.4).

IRT likewise offers much in the way of reporting. IRT can be used not only for the purpose of estimation, but also for reporting, because item difficulty and test-taker proficiency are on the same scale, which can be mobilized to add meaning to the reporting scale (Hambleton & Zenisky, 2017). Some especially powerful examples of this include the item mapping approach used by NAEP (see U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2015), the score reports used on the Massachusetts Adult Proficiency Test (Zenisky et al., 2018), and the analogous Rasch-based Wright map (Measurement Research Associates, 2010). Item mapping is conceptually connected to the work by Beaton and Allen (1992) on scale anchoring, where specific points on the reporting scale, the anchor points, are illustrated by test content.

Because testing technology has been and is ever evolving, the data that can be used for the *how* of reporting are likewise expanding. Crowdsourced learning (Milligan & Griffin, 2016) serves as but one example of the use of logstream data for scale development and, ultimately, reporting. Indeed, efforts to report data based on not just scores

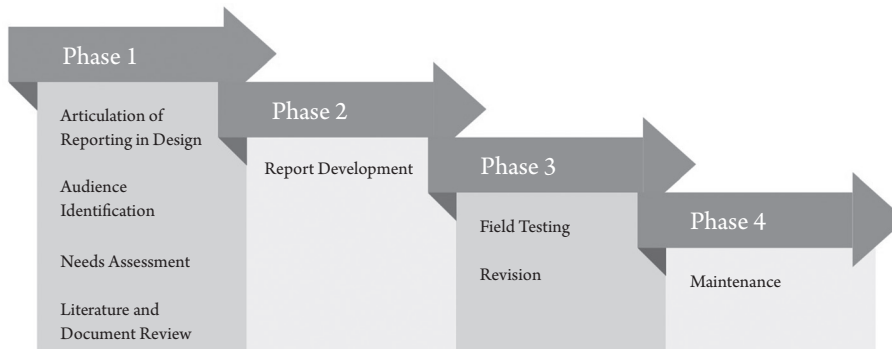
but also process data offer clear innovations for the *how* of testing (e.g., De Boeck & Scalise, 2019; Ercikan & Pellegrino, 2017; Xiao et al., 2022), as does advancing work in the field of learning analytics (e.g., Du et al., 2021; Oliva-Córdova et al., 2021; Siemens & Baker, 2012). As agencies consider scores that derive from other data sources, the meaning of scores and other results will continue to shift and reporting will accordingly need to change as well, to reflect results that are increasingly multidisciplinary and indicative of a different orientation in conceptualizing and representing performance.

Report Development Process

Quality score reporting does not happen by happy accident. It requires commitment from testing agencies to a process that is closely integrated into the larger schedule of test development activities, to define the necessary connections between the choices made in test development and the validity of the test score inferences to be communicated. (Zenisky & Hambleton, 2015, p. 601)

To this point, much of this chapter has focused on conceptual perspectives on reporting and to establish reporting as an essential activity in the tradition of the IUA of validity. With this grounding, the focus now turns to report development in earnest, and specifically models for crafting reports using thoughtful and collaborative processes. Models developed by Zapata-Rivera (2011) and Hambleton and Zenisky (2013) can be used to guide the report development process. Zapata-Rivera's (2011) model includes four main steps: "(a) gathering assessment information needs from stakeholders; (b) reconciling these needs with the available assessment information, (c) designing various score report prototypes, and (d) evaluating these score report prototypes internally and externally" (p. 37). We focus on the Hambleton and Zenisky (2013) model for the remainder of this section, but also highlight recommendations from Zapata-Rivera that apply to different stages of the development process.

Zenisky and Hambleton (2012a) introduced an initial version of the Hambleton and Zenisky model and it was described in more detail in later publications (Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015). The model was informed by knowledge of the reporting literature, best practices in test development, and direct experience with report design. Further, it stemmed from the idea that reporting efforts, like other processes including item analysis and standard setting, should follow a general cycle of development. The model was designed to be flexible and to apply to various testing contexts, so it emphasizes process considerations rather than rigid content and design requirements. As such, the model may be adapted to guide the development of reporting tools (e.g., dashboards for assessment results), as demonstrated by O'Donnell et al. (2021). Readers interested in developing such tools are encouraged to consider the Hambleton and Zenisky model, replacing "report" with "reporting tool," as well as literature from the learning analytics community on evaluating the usefulness, usability, and effectiveness of dashboards (see Corrin et al., 2019, for a review).

**FIGURE 13.9****The Hambleton and Zenisky Model for Score Report Development**

Note. Adapted from “Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design,” 2013, by R. K. Hambleton & A. L. Zenisky, in K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C., Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez, *APA Handbook of Testing and Assessment in Psychology: Vol 3. Testing and Assessment in School Psychology and Education* (pp. 479–494). American Psychological Association. <https://doi.org/10.1037/14049-023>

At a high level, the Hambleton and Zenisky model (Figure 13.9) begins with four tasks: articulating how the test, results, and reports will jointly support intended inferences (1a), identifying all groups of intended report users (1b), completing a needs assessment for each of those groups (1c), and conducting a literature and document review (1d). The next steps involve using the information obtained previously to develop prototypes (2), completing as many iterations of field testing and revisions as needed for reports to be ready for operational use (3), and implementing a maintenance plan after the reports or reporting tools have been released to the public (4).

We describe each element of the model in more detail in the following paragraphs. But, before delving into the *how* of report development, it is important to consider *who* should be involved in the process. A variety of talents and perspectives are needed to tackle reporting challenges and foster innovation. Reporting is an interdisciplinary activity, and so teams of report developers should also be interdisciplinary. As described by Slater et al. (2019), individuals on report development teams may include graphic designers, user experience practitioners, cognitive science researchers, psychometricians, assessment developers, information technology staff, and accessibility experts. Teams may include developers who serve multiple roles, and consultants may be brought in if needed. Collaboration is key in each step of the development process, as described through the lens of the Hambleton and Zenisky model.

Phase 1: Laying the Groundwork

The first element of the Hambleton and Zenisky model, *articulation of reporting in design* (1a), requires deliberate consideration of the alignment among the purpose(s) of the test, test design features, and the aims of the report(s) under development. It cannot be overstated that reporting goals should be considered early in the test development process to ensure that decisions about content, format, and test length are in agreement with the kinds of information an assessment should provide. If there are distinct claims

about reports and score interpretations that are specified at the beginning of the test development process, then reports that are intended to support the IUA should likewise be framed out at that same time to ensure that test design choices actually gather and provide the data to support the eventual reporting activities and report contents, in effect a backward mapping of reports to the test development process. At this initial stage, it may be helpful to begin preparing a “prospective report”—a simple mock-up of what the final report might include that can be refined as more information becomes available and used to inform test development discussions (Zapata-Rivera et al., 2012).

The next element, *audience identification* (1b), involves naming the user groups that will need information from the report(s) to draw conclusions or take specific actions. In the K–12 context, the target users for a test taker–level report may be students and parents, while teachers and administrators may be the target users for a group-level report. In the credentialing context, those candidates scoring above the minimum passing standards and those below the minimum passing standard may perhaps be considered distinct user groups, because their reporting interests may be quite different depending on what side of the standard they are on. Jaeger’s (2003) work illustrated the multiple user groups that may rely on reports from an assessment program—up to nine groups in the case of U.S.-based NAEP. It must be understood as well that in the process of identifying user groups, such groups are composed of individuals, who bring varying interests, background knowledge, and needs to assessment data communication. In this step, audiences must be differentiated, but the variation within each user group must also be acknowledged.

Thus, after identifying the target audience, a *needs assessment* (1c) should be conducted to understand each user group’s information needs, as well as characteristics that might influence how members of the group approach assessment results. From an audience analysis perspective, Zapata-Rivera and Katz (2014) recommended focusing on the following areas:

- Needs: Users’ purpose for reviewing a report or reporting tool, exemplified by questions such as, “What inferences about or decisions concerning test takers does the user want to make?” (p. 447)
- Knowledge: What users already know and what information they need to fully comprehend the report, exemplified by questions such as, “What knowledge gaps might interfere with correct interpretation of the score report?” (p. 447)
- Attitudes: The feelings or biases that might influence users’ interpretations of the information reported, exemplified by questions such as, “What do users think about assessment generally?” and “What are their expectations about the test takers?” (p. 447)

The needs assessment should be viewed as a broad-based inquiry, extending well beyond a simple investigation of user groups’ familiarity with statistical terms (Zapata-Rivera & Katz, 2014; Zenisky & Hambleton, 2015). It must take place early in the test development process, to inform test development decisions and be informed by user reporting

needs and interests. Zenisky and Hambleton (2015) also recommended gathering general feedback on how users would like to access the report and what data elements they would like to see, which can then be reconciled with what the assessment program is able to support. Report developers may collect information for this step directly through focus groups or surveys or indirectly by reflecting on their own experience with the target audience. If an early prospective report was created, it can be shared with stakeholders, if appropriate, and refined based on key feedback from this stage.

The last element in Phase 1 of the Hambleton and Zenisky model is a *literature and document review* (1d). Report developers should be familiar with general principles from the literature (e.g., Hattie, 2010; Slater et al., 2019; Zenisky & Hambleton, 2012a), as well as findings specific to report components under consideration, such as visual displays (e.g., Hegarty, 2019; Ryan, 2006; Wainer, 1997b), language (Roduta Roberts et al., 2018), or representations of measurement error (e.g., Kannan et al., 2018; Zapata-Rivera et al., 2019; Zwick et al., 2014). It is also helpful to note which methods have been employed to gather feedback on reports, paying special attention to those involving similar assessments and/or audiences.

A number of organizations such as state departments of education to credentialing agencies regularly post sample reports and interpretive guides online (Knupp & Ansley, 2008). To learn from current design approaches, it is useful to examine those samples (or reporting tools, if applicable), keeping track of content and design choices of interest based on the assessment data available and audience characteristics. There is no combination of content and design elements that guarantees universal success, but reviewing a series of complete reports and interpretive guides helps the report development team in the process of beginning to imagine the possibilities (Goodman & Hambleton, 2004).

Indeed, interpretive guides play a pivotal role for report users seeking additional guidance. Reports should be self-contained and concise, including all essential information to support accurate inferences. However, paper or online interpretive guides should still be available for users who may not be as familiar with the context of the assessment or who may be interested in learning more. Goodman and Hambleton (2004) identified several interpretive guide components for consideration, including answers to common questions about the assessment, more information about its content and purpose, suggestions to improve performance, and guidance on where to find additional resources.

Phase 2: Report Development

In Phase 2, *report development*, the information gathered earlier in the process should be used to create a set of prototypes that are “aligned with the test’s goal(s), audience-specific and also rooted in best practices” (Zenisky & Hambleton, 2015, p. 593). Crafting multiple prototypes is helpful so that members of the target audience can react to a range of content and design options. Although it is not impossible, it seems unlikely that a group of report developers would get all elements right on the first and only prototype developed.

The prototypes may have text and design elements in common, and it is up to the report developers to decide how detailed the initial set should be. To the extent possible, the prototypes should reflect the “look and feel” of a complete report or reporting tool and be informed by data gathered in Phase 1. Prototypes missing key features may lead to inaccurate results during field testing, such as members of the target audience indicating that a design is not cluttered only because some or all text is missing, or having trouble understanding the flow of information because a major graph or other element has not yet been added. If multiple reports are needed, a list showing the audience and core data elements for each can assist with tracking development progress.

Phase 3: Field Testing

Always field-test graphs, figures, and tables on focus groups representing the intended audiences; many important things can be learned from field-testing report forms such as features of reports which may be confusing to readers. (Hambleton & Slater, 1997, p. 18)

Phase 3 involves an iterative cycle of *field testing* and *revisions* until the report is ready for implementation. An internal review by colleagues from different backgrounds or with sufficient knowledge of the target audience may be conducted prior to external review. Table 13.1 offers a comprehensive list of evaluation questions that may be used at this stage. The questions are grounded in best practices from the literature and are meant to promote critical reflection. Following internal review, recommendations that lead to minor changes can be readily implemented (e.g., wording suggestions to improve

Table 13.1. Updated Evaluation Form for Reviewing Reports

Report Element	Report Review Questions
I. Overall	A. What are the key intended interpretations of the report? B. How does the report reflect the interests and informational needs of key stakeholders? C. In what specific ways does the report present information that aligns to the purpose(s) of the assessment? D. What evidence has been gathered to support the validity of the report for intended user groups?
II. Content— report intro- duction and description	A. Does the report have a title clearly and descriptively identifying whether it is for an individual or a group? B. For group reports, how are the parameters for inclusion in the group defined? C. How are details provided about the content of the test(s) being reported, if it is not common knowledge for the audience? D. How is the purpose(s) of the test expressed, given the intended audience? E. If present, in what way does the introductory statement set a positive tone for the report?

Report Element	Report Review Questions
III. Content— scores and performance levels	A. How many different reporting scales are used on the report? Is the meaning and range of each score scale communicated, and how? B. How are the performance categories or psychological states being used (e.g., <i>Basic</i>, <i>Proficient</i>, <i>Advanced</i>) reflective of intended inferences, not stigmatizing, and described sufficiently for the intended audience? C. If it is not obvious, how is information provided to guide interpretation and use for each score and classification? D. What strategies are used to guide users away from known misinterpretations and misuses? E. How is the topic of score imprecision handled, to promote interpretation, for each result reported (i.e., overall and subdomains)? F. Have “probabilities” or “conditional probabilities” been used? If they are used, is the explanation clear? G. Is there sufficient information for the reader, without being overwhelming?
IV. Content— other perfor- mance indica- tors	A. Is there any linking of test results to possible follow-up activities? For example, with educational tests, are the results linked to possible instructional follow-up? B. If present, are relevant reference group comparisons reported with information on appropriate interpretations? C. If norms are provided, what steps are taken to describe the reference group in sufficient detail? D. If present, how are results of performance on individual test questions reported to facilitate use and understanding? E. If present, how are reports of scores from other recent and relevant tests explained?
V. Content—other	A. How does the report provide information about where to direct questions? B. Does the report provide links to additional resources about the test, testing program, and/or understanding test-taker performance?
VI. Language	A. What steps have been taken to ensure that the report is free of statistical and other technical jargon and symbols that do not facilitate or promote understanding and interpretation? B. What steps have been taken to ensure that the text is clearly written for users? C. If footnotes are used, are they clearly written for the reader? D. If the report (or ancillary materials) is translated/adapted into other languages, how is the translation/adaptation carried out? What steps were taken to validate the translated/adapted version?
VII. Design	A. What visual and/or narrative strategies are used to highlight the information that is most important based on the purpose(s) of the assessment? B. How is the report clearly and logically divided into distinct sections to facilitate readability?

(continued)

Table 13.1 (continued)

Report Element	Report Review Questions
	C. What visual or narrative strategies are used to communicate the key score information? D. Is the font size in the different sections suitable for the intended audience? E. What steps have been taken to ensure that the graphics (if any) are presented clearly to the intended audience? F. Is there a mix of text, tables, and graphics to support and facilitate understanding of the report data and information? G. What evidence has been collected to suggest that the report looks friendly and attractive to users? H. What are the steps taken to ensure that the report has a modern “feel” to it, with effective use of color and density (a good ratio between content and white space)? I. What steps have been taken to ensure that the report is free of irrelevant material and/or material that may not be necessary to address the purposes of the report? J. What evidence is gathered to suggest that the “flow” for reading the report is clear to the intended audience? K. How does the report align in layout and design to related materials published by the testing program?
VIII. Interpretive guides and ancillary materials	A. Is there an interpretive guide prepared, and if so, what steps have been taken to ensure that it is informative and clearly written? Has it been field tested? Are multiple language versions available to meet the needs of intended readers? B. If there is an interpretive guide, is there an explanation of both acceptable and unacceptable interpretations of the test results?

Note. Adapted from “Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design,” 2013, by R. K. Hambleton & A. L. Zenisky, in K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C., Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez, *APA Handbook of Testing and Assessment in Psychology: Vol 3. Testing and Assessment in School Psychology and Education* (pp. 479–494). American Psychological Association.
<https://doi.org/10.1037/14049-023>

clarity, small formatting changes), but more complex requests should be considered with caution, and it may be best to wait for external review to confirm the request is aligned with report users’ wants and needs.

During external review, it is imperative that report developers choose field-testing approaches that provide information not only on preferences, but also on comprehension. The report elements that users like and the elements they understand are not always the same (e.g., Wainer et al., 1999). Different approaches yield different information, and accordingly, it may be advantageous to use more than one method to obtain data about reports using multiple approaches.

Surveys may be an efficient way to gather data on both areas and reach a high number of participants. Typically, surveys used for field testing include images of the prototypes and a mix of rating scale and open-ended questions. Participants may be more willing to provide candid answers and may enjoy being able to explore

materials and answer questions at their own pace. However, depending on their level of motivation and interest, engagement with the questions may be lower than for methods with active facilitation. Last, report developers must decide whether to make the survey anonymous or request contact information to allow for follow-up questions.

Another option is to conduct focus groups (e.g., Forte Fast & Tucker, 2001; Ryan, 2006; Zenisky, Delton, & Hambleton, 2006; Zenisky, Hambleton, & Smith, 2006). They offer the possibility of hearing correct interpretations or misinterpretation of reports, allow asking follow-up questions as needed, and can be useful for getting answers to prepared questions as well as exploring unplanned but important topics that arise during discussion. To encourage full participation, Zenisky and Hambleton (2015) recommended holding separate focus groups with members of different user groups and being mindful of potential power imbalances among participants. Probing comprehension in a group setting can be challenging; participants may be unwilling to venture interpretations or admit that something is difficult to understand. An alternative to asking direct comprehension questions is asking questions such as:

- Do you think other [parents/teachers/students] would understand this information?
- Is there something in particular that might be confusing to them?
- What is your interpretation of this report?
- What would you do next now that you have interpreted this report?

Such questions can lead to many interactions among focus group members that can inform the adequacy of the report and areas for improvement.

An additional source of evaluative questions in this regard can be sourced from Ryan (2006). These can be tailored to be relevant to the intended user group as well as the person doing the ratings.

- Will a [user] find this information helpful?
- How could a [user] use this information?
- Could this information be modified to be more informative or helpful?
- How can this information be best presented?
- Might there be any problems in how this information is used?

These particular questions were posed as conceptual points to help study participants in the context of rating different proposed reporting formats, but can also be viewed as stand-alone questions that could be asked of participants.

Interviews are the most time-intensive option, but can provide rich information on an individual's perceptions, and participants may feel more comfortable sharing honest feedback on prototype elements that are difficult to interpret in this individual setting. Like focus groups, interviews can be based on a set of prepared questions or follow a more conversational style. Unlike focus groups, they can incorporate techniques

such as think-aloud protocols, where users are asked to voice their thought processes and reactions as they interact with a prototype. When there is particular interest in understanding how users navigate prototypes, eye tracking and direct observation may also be used.

Regardless of methods used to gather data on reports, report developers must be acutely involved in the process of identifying participants for feedback on reports and seek out research participants who may be able to provide wide-ranging perspectives on report prototypes. It is not sufficient to name intended user groups at a general level (e.g., “parents,” “teachers”) while ignoring the subgroups that are present within such groups. Broadly speaking, sex, race/ethnicity, socioeconomic status, proficiency in the primary language of the report, and technology familiarity (for web-based reporting efforts) may provide initial sampling frames for recruiting participants. For example, in a reporting jurisdiction with a high proportion of Spanish-speaking families, it would be appropriate to develop a sampling plan that ensures representation across the diversity of language status in that jurisdiction. This approach not only benefits a specific report document but also can glean insight about communication and access strategies around reporting. Similarly, if a report is developed in certification and licensure contexts that offer a diagnostic component for failing/low-performing candidates, with the aim of providing actionable information, then failing/feedback from low-performing candidates would be something that would likely be useful to include on a sampling plan. The process of gathering feedback at any point in the report development cycle can be time-consuming and complicated, but careful consideration of the research participants, including clear articulation of relevant subgroups for any individual user group and outreach to engage those subgroups, cannot be overlooked as an aspect of validating reports. Just as test developers seek out diversity in test specifications panels, item writers, and standard-setting panels, so too must diversity be prioritized in participatory research around reporting.

After analyzing the results from field testing and identifying areas of consensus, report developers must decide which content and design changes to make. Some improvements may be obvious upon reviewing the feedback, while others may not be immediately feasible or take additional consideration. For example, there may be consensus that a particular layout is confusing, but no suggestions on how to improve it. Depending on the magnitude of the revisions, additional field testing may be needed to investigate whether the changes had the intended effect. This phase is intended to be iterative, when appropriate.

Phase 4: Maintenance

The *maintenance* phase begins once reports become operational. As explained in the validity section, investigating whether reports are interpreted and used as intended is critical. This phase may include periodic evaluations using some of the same methods employed for field testing in addition to continuous monitoring of inquiries and comments from

report users received via customer support channels, online forums, and other outlets. If content or design flaws are identified, “easy fixes” may be implemented right away, or potential revision ideas may be accumulated over a certain period to implement a number of changes at once. As with other phases of the model, there are no strict rules about how often to conduct maintenance research and what methods may be involved, but critically there should be a plan in place, and the plan should be aligned with the process for gathering evidence of actual interpretation and uses of the information reported.

Summary

The Hambleton and Zenisky model presented here, supported by the principles and ideas from Zapata-Rivera (2011), is an approach to report development that aligns with research and practice. The driving force behind this model is to ensure that results reports, as a public-facing end product of the test development process, meet the known and articulated needs of the various intended users, through a collaborative and iterative series of steps. Evaluation is a critical aspect of this work, and the checklist provided in Table 13.1 offers report development teams opportunities to reflect on their work and the final report to demonstrate its validity relative to the test’s purpose and the communication process.

WHAT WE HAVE LEARNED

A theme throughout this chapter is that developing effective reports, whether static or interactive, requires careful consideration of the who, what, where, when, why, and how: specifically, who will use the results, what data are available, where the results will be reported, when they will be released, why results are needed, and how data will be turned into results. These factors are interdependent and, combined, create numerous possibilities. Teams developing reports or reporting tools may avail themselves of models such as the one in the previous section, but because effective reporting is so context specific, summaries of findings by report element may be difficult to find.

Gotch and Roduta Roberts (2018) conducted a review of 60 studies on individual reports published between 2005 and 2015, summarizing areas of focus, theoretical frameworks of communication, and data characteristics rather than findings regarding specific report elements. They noted that “data sets were often small or localized to a single context” (p. 46), highlighting the issue of generalizability. Although there have been no meta-analyses of findings from studies on results reporting to date, likely due to how context dependent such findings tend to be, some general guidance is available.

To this end, Figures 13.10 through 13.17 offer key recommendations from five sources: Goodman and Hambleton’s (2004) review of results reports and interpretive guides; Ryan’s (2006) *Handbook of Test Development* chapter on reporting practices, issues, and trends; Hattie’s (2009) principles for promoting the validity of reports; Zenisky and Hambleton’s (2012b) guidelines for developing online reporting resources; and Slater et al.’s (2019) chapter on designing reports for large-scale testing programs.

These recommendations are grouped on the basis of eight dimensions of reports and reporting: overall; content—report introduction and description; content—scores and performance levels; content—other performance indicators; content—other; language; design; and interpretive guidance and ancillary materials (these dimensions align to the report elements of the checklist presented in Table 13.1).

These works offer advice based on findings and practical experience from scholars who have dedicated substantial portions of their careers to improving the communication of assessment results. Although visualization ideas from Wainer are not directly represented, Wainer's extensive contributions and publications significantly informed the work of Goodman and Hambleton (2004) and Ryan (2006).



Goodman & Hambleton (2004)

"Consideration should be given to the creation of specially designed reports that cater to the particular needs of different users." (p. 219)

"Personalize the student score reports and interpretive guides." (p. 219)

"Reports should be piloted with members of the intended audience." (p. 219)



Ryan (2006)

"A score report should be related to content standards as clearly and explicitly as possible." (p. 705)



Hattie (2009)

"The validity of reports is a function of the reader's correct and appropriate inferences and/or actions about the test takers' performance based on the scores from the test." (p. 3)

"Evidence is needed to demonstrate how readers are interpreting reports." (p. 4)

"A report should be timely to the decisions being made (formative, diagnostic, summative, and ascriptive)." (p. 13)

"Anchor the [report or reporting] tool in the task domain." (p. 9)



Zenisky & Hambleton (2012a)

Static: "Pay attention to the reporting interests and needs of users, as presenting results in static format means users see only what is shown to them." (p. 179)

Interactive: "Have users articulate their data analysis needs." (p. 181)

"Look into what others have done. Many states and publishers are being very creative in developing online interactive tools." (p. 181)



Slater et al. (2019)

"Design the report so that viewers can see and understand the most important information in 10 seconds or less." (p. 98)

"Make sure that the report design can accommodate unusual but possible conditions, such as very long names or very low or high scores." (p. 99)

"Follow Web Content Accessibility Guidelines (WCAG; Caldwell, Cooper, Reid, & Vanderheiden, 2008)." (p. 99)

FIGURE 13.10
Key Reporting Recommendations: Overall



Goodman & Hambleton (2004)

"Include an easy-to-read narrative summary of the student's results at the beginning of the student score report." (p. 219)

"Devices such as boxes and graphics should be used to highlight main findings." (p. 219)



Ryan (2006)

"Score reports should highlight important results in some way, (e.g., boxes, boldface type)." (p. 706)



Hattie (2009)

"A report should provide justification of the test for the specific applied purpose and for the utility of the test in the applied setting." (p. 11)



Slater et al. (2019)

"Emphasize the most important information in the score report. . . . The most important parts of the report should command the most attention." (p. 98)

FIGURE 13.11

Key Reporting Recommendations: Content—Report Introduction and Description



Goodman & Hambleton (2004)

"Include all information essential to proper interpretation of assessment results in student score reports." (p. 219)



Ryan (2006)

"[Results] should be reported in relation to performance standards." (p. 705)

"[Results] should be reported at the finest level of detail for which reliable information can be provided." (p. 706)

"A score report should include information about precision for all scores presented." (p. 706)



Hattie (2009)

"Maximize interpretations and minimize the use of numbers." (p. 5)



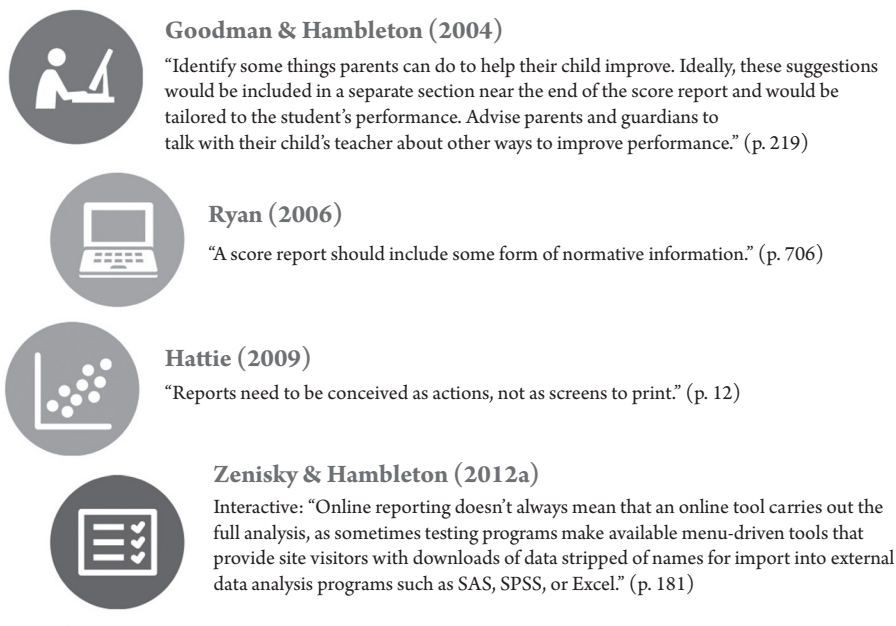
Zenisky & Hambleton (2012a)

Static: "Try out different report formats (summary, highlights, full-length reports) and data displays (text, graph, and tables) with intended audiences to ensure that materials are understood and the conclusions being drawn are appropriate." (p. 179)

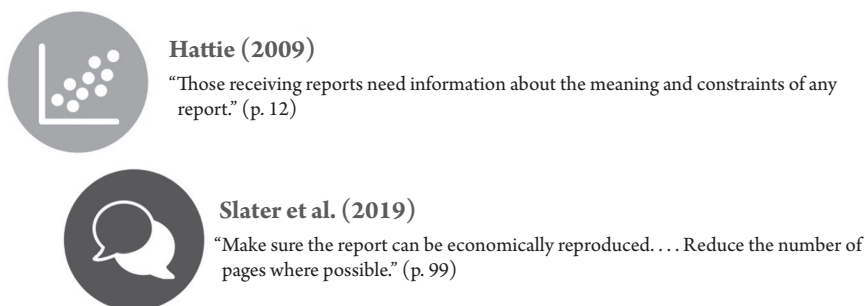
Interactive: "Start with things that are manageable, and look into developing comparatively simple web interfaces supported by databases that let users select results for a content area, unit of analysis (school, city/town, region/county, state), and perhaps limited demographic characteristics." (p. 181)

FIGURE 13.12

Key Reporting Recommendations: Content—Scores and Performance Levels

**FIGURE 13.13**

Key Reporting Recommendations: Content—Other Performance Indicators

**FIGURE 13.14**

Key Reporting Recommendations: Content—Other

The Road Ahead: Opportunities in an Evolving Reporting Landscape

While much research and effort has been exerted to formalize reporting efforts within the broader domain of test development, the ongoing reimaginings of assessment spurred in part by technological advances within and outside assessment demonstrate that reporting is a topic that itself is not static, and work in this area is ongoing. In this section, several priority areas for future work are briefly highlighted.

**Goodman & Hambleton (2004)**

"Score reports should include easy-to-read text that supports and improves the interpretation of charts and tables." (p. 219)

"Small fonts, footnotes, and statistical jargon should be avoided." (p. 219)

"Key terms should be defined, preferably within a glossary." (p. 219)

**Ryan (2006)**

"Avoid jargon unfamiliar to the intended audience." (p. 706)

"Provide an explanation or glossary for any measurement terms used." (p. 706)

"Use text to explain graphs, charts and tables." (p. 706)

**Slater et al. (2019)**

"Ensure that any language in the report is appropriate to its intended audience, at the proper reading level." (p. 98)

"Avoid technical language that might be difficult for non-experts to understand." (p. 98)

FIGURE 13.15**Key Reporting Recommendations: Language****Goodman & Hambleton (2004)**

"Score reports should be clear, concise, and visually attractive." (p. 219)

"Care should be taken to not try to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes)." (p. 219)

"Data should be grouped in meaningful ways." (p. 219)

**Ryan (2006)**

"Score reports should be clear, as simple as possible, and uncluttered." (p. 706)

"Use simple and clear graphs, charts, and tables." (p. 706)

"Score reports should use print features such as font size, style, and spacing that make it as easy as possible for the reader to understand the report." (p. 706)

**Hattie (2009)**

"Each report needs to have a major theme." (p. 7)

"A report should be designed to address specific questions." (p. 11)

"Readers of reports need a guarantee of safe passage...and destination recovery." (p. 4)

"The answer is never more than 7 plus or minus two." (p. 6)

"A report should minimize scrolling, be uncluttered, and maximize the 'seen' over the 'read.'" (p. 10)

**Zenisky & Hambleton (2012a)**

Interactive: "Different users have different capabilities for understanding quantitative data, so look into presenting data in multiple formats when possible (tables, graphs, narrative, etc.);" (p. 181)

**Slater et al. (2019)**

"Create a strong visual hierarchy that guides the viewer's eye appropriately through the report" (p. 98)

"Eliminate visual clutter . . . Avoid repeating elements of the report, except where repeating them makes the report clearer." (p. 98)

"Avoid using lines that add to visual complexity. Instead, use shaded areas to delineate space." (p. 98)

"Use visual embellishments such as icons only when they make it easier for users to correctly interpret the report. . . Avoid any visual element that may have a negative connotation." (p. 98)

"Colors, used meaningfully, can make the report easier to interpret. However, make sure that the report can convey all its information even if the viewer cannot accurately differentiate colors or if the report is printed or photocopied in black and white." (p. 99)

FIGURE 13.16**Key Reporting Recommendations: Design**



Goodman & Hambleton (2004)

"Include detailed information about the assessment and score results in a separate interpretive guide." (p. 219)

"Include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level." (p. 220)

"Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports." (p. 220)



Zenisky & Hambleton (2012a)

Informational Web Pages:

"Ask stakeholders what information about the testing program would be helpful for them to know." (p. 182)

"Review other testing programs' websites to find out the kind of programmatic documentation they make available." (p. 182)

"Post the technical manual and updates online as available." (p. 182)

FIGURE 13.17

Key Reporting Recommendations: Interpretive Guides and Ancillary Materials

Reporting in the Formative Assessment Context

Reporting results from formative assessments is a growing area of research and operational development. While the focus of most reports for summative interpretations is a total score indicating overall achievement in relation to a target construct, such a broad indicator of performance is typically much less interesting and useful in formative contexts. Formative interpretations use a range of formats and delivery methods, but their defining characteristics are being administered *during* instruction and providing feedback to inform adjustments to teaching and/or studying approaches. Informing immediate action to support student learning is central to formative assessment, creating unique challenges and opportunities for reporting.

While expediency is desirable in all testing contexts, it is essential to formative interpretations: results must be ready in time to make a difference. Rapid reporting increases the relevance of the information provided and the probability that educators will act on the results (Brown et al., 2019). Specificity is also important. Reports for formative interpretations should provide sufficient detail so that users can identify areas of strength and weakness and know where to focus to make improvements. These finer grained results have traditionally included item-level feedback and scores or achievement levels for specific content areas or attributes. More recently, the options have expanded to include process data.

Process data may include time spent on each assessment task, logs of which resources were used at different points, logs of answer changes, click streams, and keystroke logs. Turning such data into insights holds substantial promise, but perhaps most especially in formative contexts, because it can paint a richer picture of student learning. An example of this is found in Milligan & Griffin (2016), who elaborated a generalized six-step methodology for constructing measures using click-stream data. However, research on how to do this, how to report process data, and how to establish the validity of this practice is still in its infancy, and there is much yet to be learned (Provasnik,

2021; Zapata-Rivera et al., 2021). Analysis approaches drawn from the cognitive diagnosis modeling literature, such as the attribute hierarchy method, may also be helpful here (Roberts & Gierl, 2010).

With or without process data, formative assessment results tend to be more detailed than results for other types of assessments, creating the challenge of designing reports or reporting tools that communicate detailed, personalized feedback in ways that are clear and actionable in addition to being automated to support rapid reporting. This may be done by combining digital reporting with assessment designs that provide finer grained results (e.g., Brown et al., 2019) and employing specific strategies such as the use of formative hypotheses to offer suggestions on next steps (Zapata-Rivera et al., 2012) and online tutorials to support comprehension of the reporting tool (Brown et al., 2019). The benefits of feedback to teaching and learning are well documented, and so are the complexities of providing feedback effectively (Hattie & Timperley, 2007; Wisniewski et al., 2020). Technological advances are creating exciting opportunities for advancing the ways we report feedback from formative assessments.

Leveraging Technology for Reporting

Interest in using technology to better communicate assessment results has steadily grown since the start of the millennium. At the time of Goodman and Hambleton's (2004) review of K–12 reporting across 11 U.S. states, only one state posted information related to results reports online. A few years later, this practice had become so common that it prompted the development of guidelines for informational web pages about results reports as well as guidelines for online results-oriented documents and interactive tools (Zenisky & Hambleton, 2012b). Bulut et al. (2020) also proposed recommendations for providing feedback from assessments in an online environment, and Hattie's (2010) commonly cited principles for report development arose from the context of developing web-based, interactive reporting tools.

Many of these recommendations and principles are similar to those for paper-based reports. For instance, Zenisky and Hambleton (2012b) suggested “presenting data in multiple formats when possible” to support users with different information processing needs (p. 181), Hattie (2010) advised that “each report needs to have a major theme” (p. 7), and Bulut et al. (2020) recommended “an aesthetically pleasing design without information overload” (p. 64). Bulut et al.'s (2020) remaining recommendations, presented verbatim, echo several of the guidelines from Figures 13.10 through 13.17.

- The score report should be tailored to meet the needs and characteristics of the target audience, such as students, parents, and teachers (Hambleton & Zenisky, 2013; Zapata-Rivera & Katz, 2014).
- The score report should present the feedback in different forms, including narrative text, tables, and figures.
- The layout of the score report should be simple, with key results highlighted (Goodman & Hambleton, 2004; Slater et al., 2019).

- Feedback presented in the score report should include a set of actions that students can take to improve their future performance (Daniels & Bulut, 2019; Hattie, 2009; Jonsson, 2012).
- If interactive elements (e.g., visuals and tables) are to be used, how students will interact with these elements should be considered in the design process (Bulut et al., 2020; Slater et al., 2019).
- Usability studies with students should be carried out to test whether the content of feedback is easy to follow (Slater et al., 2019; Zenisky & Hambleton, 2012b).

What is different in technology-based reporting is that there is usually no physical constraint on how many elements can be included, and interactivity offers possibilities that were not available before, such as on-demand interpretive text, performance details, and interactive tutorials (e.g., Zapata-Rivera et al., 2016).

In thinking of how to characterize next-generation reports in other innovative ways, it may be that the near future may be app based and use the screen and app-based functions to engage users in ways that current results reports may not (Linares-Vásquez et al., 2017). In the very near future, what users think of as reports may look and feel very different from current “score reports.” Other forward-looking research in the areas of big data and visualization may be similarly instructive, in terms of integrating machine learning, developing novel visualizations, conceptualizing different patterns of interaction with data, and alternative user interfaces (Andrienko et al., 2020). This thinking in the area of big data echoes the model-based approach advanced by Zenisky and Hambleton, 2012a), since Andrienko et al. (2020) suggested understanding and “designing the user interactions first” (p. 4), followed by development of systems to support such interactions.

With so many content and design options available, report developers must be very familiar with the needs and preferences of their audience to develop interfaces that will seamlessly guide them—interacting with online resources and tools to explore score data and other results successfully involves greater user engagement than reading a static report. It also may require background knowledge about the assessment itself and the construct measured to formulate specific questions and use the tool or tools as intended. It is also helpful to consider likely misinterpretations and misuses, which become more numerous when interactivity is added, and make adjustments to support the purpose of the assessment. These are critical considerations that developers of online reporting tools must carefully understand and articulate at the outset of that development work, to ensure that these issues are addressed throughout report tool development, including attention paid to providing context and background knowledge for use of any tool. Then, several rounds of gathering user feedback and making refinements are usually needed when developing interactive tools, and relying on a model such as the one proposed by Hambleton and Zenisky (2013) may aid in this process.

Some challenges are unique to online reporting, such as matters of access and ease of navigation. The way users access online reports or reporting tools should be secure

and convenient. Further, the document or tool itself should be easy to navigate, providing users a “guarantee of safe passage” (Hattie, 2010, p. 4), with the most important information made prominent by the design and visual cues, helping users decide where to look next. If the practices of online reporting since the year 2000 are any indication, it is highly likely that online reporting will continue to grow in popularity, bringing about new ways to communicate information from assessments and interesting challenges.

Data Visualization

Approaches to visualizing assessment results have continued to evolve. Classic resources on displaying quantitative data remain relevant (e.g., Kosslyn, 2006; Tufte, 1983, 1990; Wainer, 1997a, 2005) and can now be supplemented by more recent findings from fields such as cognitive science, information visualization, and user experience. Hegarty (2019) provided a thorough discussion of contemporary data visualization best practices and principles that apply specifically to results reporting.

Additionally, the refinement and greater availability of data visualization tools has allowed psychometricians to become more involved in prototyping and producing graphics for reports and reporting tools. Consider the functionality of the following selection of popular open-source data visualization packages for the programming language R (similar tools are also available for Python):

- *ggplot2* (Wickham, 2016): A package for creating static graphics with elegant default features and flexible customization options, based on Wilkinson’s (2005) *Grammar of Graphics* structured approach for constructing visualizations (<https://ggplot2.tidyverse.org/>)
- *plotly* (Sievert, 2020): A package for creating modern, interactive graphics powered by a JavaScript charting library; it may also be used to make *ggplot2* graphics interactive using the *ggplotly* function (<https://plotly-r.com/>)
- *shiny* (RStudio, 2020): A package and web application framework for creating interactive dashboards with R, often used in conjunction with *plotly* (<https://shiny.rstudio.com/articles/#deployment>)

Because results reports are composed of both content and communication, hand in hand with these tools for visualization is an evolution in what reports look like and how the data are accessed by stakeholders. One increasingly common strategy for reporting is the use of digital databases and user-selected reporting queries, and indeed this has shifted the control of the data story from psychometricians and report developers to users (Zenisky & Hambleton, 2012b).

Along the lines of strategies for visualization, a next frontier for reporting, especially with respect to next-generation assessments, is to continue to build on interdisciplinary contributions from cognitive science, marketing and communications, and dissemination practices in other scientific domains to leverage existing and emerging strategies for data understanding and use. In particular, the work of Hegarty (2011,

2019; Padilla et al., 2018), among others (e.g., Hullman et al., 2011; Ratwani et al., 2008), has helped bridge understandings of visual–spatial display cognitions to guide action in a number of areas, such as radiology reports (Alarifi et al., 2021), clinical monitoring (Khasnabish et al., 2020; Reese et al., 2020), meteorology and climate (Argyle et al., 2017; Gerst et al., 2019; Harold et al., 2016), and mapping (Hegarty et al., 2016; Johannsen et al., 2018). There is much to be gained by leveraging modern visualization tools and insights from other fields to improve the visual elements of reports and reporting tools.

Reporting for Next-Generation Assessments

Assessment is changing. From constructs measured to delivery mode, from item formats to sequencing of measurement opportunities in adaptive and gamified assessments, many tests today are substantively different in one or more significant ways than those of just a decade or two ago. As this evolution in tests and measurement continues, it is at least in part driven by users and stakeholders who seek not only different information from assessments but also to use test data more efficiently and effectively (Brown et al., 2023). In this way, reporting—beyond a global score or performance level—is centrally ingrained in the next generation of assessments.

Next-generation assessments are not marked first by technology, but rather a more fundamental shift in paradigm to prioritize process and information, rather than status or outcome. This will necessitate continued work not only on reporting but also next-generation psychometrics, to support the inferences and information that are increasingly sought. The ongoing efforts to report subscores reliably are but a precursor to the challenges that lie ahead in terms of making useful sense of process data (e.g., Bergner & von Davier, 2019; Provasnik, 2021) and, importantly, to devise ways to communicate these data in ways that are supported by evidence. Emerging work in areas such as big data, natural language processing, and digital learning environments will be critical to informing these desired shifts in report contents (Cope & Kalantzis, 2015; Mislevy et al., 2012).

CONCLUSION

Communicating test score information matters. (Hambleton & Zenisky, 2013, p. 492)

Reporting assessment results is not easy. It is the culmination of a long process of assessment design and development, fundamentally marked by a systematic and serious effort to provide information about individual and/or group performance relative to educational and psychological constructs of interest. As discussed here, reports can be conceptualized through the lens of a data story, where the elements of the report are purposefully chosen, arranged, and field tested with prospective users to communicate specific information about test performance. Much research has focused on reporting in recent years, and the addition of guidance to formalize the process (Hambleton &

Zenisky, 2013) gives structure and purpose to activities that are critical for the development of useful report documents and tools.

In this chapter, the intent has been to reinforce the importance of reporting as part and parcel of the IUA for an assessment. If users are to interpret and use any assessment, they must have access to data in a way that supports those efforts. The work of Kane (2006, 2015) and Hattie (2009, 2010) is instrumental in shaping the perspective on validity and reporting we adopted, and the idea of “interpretability” as suggested by O’Leary et al. (2017) adds an interesting and potentially useful dimension to the conversation around validity and the role of reporting in supporting test interpretation and use.

In some ways, it may almost be easier to focus the conversation about report development on graphics, colors, layout, and design because those are easily manipulated key features of the end product of the report development process. However, as is evident in the earlier conversation around effectiveness, the true test of the success of a report is in its use. The criticality of engaging in reporting needs and wants at the outset of test development cannot be overstated, because the priorities and decisions made early on define the parameters of the data (from a validity perspective) that will be reported later. Interpretations that may be informally desired but are not discussed, defined, and/or prioritized from the beginning are likely to be *unvalidated* and, therefore, relegated to inappropriate uses of results.

If there is a final point to be made, it is the importance of when the conversation about interpretation and use—and reporting—takes place, so that test developers and stakeholder groups alike have a clear understanding of the *what* of reporting, to ensure that the assessment fulfills those known reporting needs (and wants). This takes time, it takes a team, and it takes planning. The evidence for effectiveness (understood as use) comes later, through Phase 4 of the Hambleton and Zenisky model (2012a, 2013) with evaluation and maintenance, but there is no substitute for a strong foundational process for reporting, building on needs assessment, stakeholder input, and consideration of reporting strategies, always, always, in light of test purpose and use.

REFERENCES

- Alarifi, M., Patrick, T., Jabour, A., Wu, M., & Luo, J. (2021). Designing a consumer-friendly radiology report using a patient-centered approach. *Journal of Digital Imaging*, 34(3), 1–12. <https://doi.org/10.1007/s10278-021-00448-z>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Andrienko, G., Andrienko, N., Drucker, S., Fekete, J. D., Fisher, D., Idreos, S., Kraska, T., Li, G., Ma, K.-L., Mackinlay, J., Oulasvirta, A., Schreck, T., Schumann, H., Stonebreaker, M., Auber, D., Bikakis, N., Chrysanthis, P., Papastefanatos, G., & Sharaf, M. (2020). Big data visualization and analytics: Future research challenges and emerging applications. CEUR Workshop Proceedings, 2578. <http://ceur-ws.org/Vol-2578/BigVis1.pdf>

- Argyle, E. M., Gourley, J. J., Flamig, Z. L., Hansen, T., & Manross, K. (2017). Toward a user-centered design of a weather forecasting decision-support tool. *Bulletin of the American Meteorological Society*, 98(2), 373–382. <https://doi.org/10.1175/BAMS-D-16-0031.1>
- Bach, B., Stefaner, D., Boy, J., Drucker, S., Bartram, L., Wood, J., Ciuccarelli, P., Engelhardt, Y., Köppen, U., & Tversky, B. (2018). Narrative design patterns for data-driven storytelling. In N. Riche, C. Hurter, N. Diakopoulos, & S. Carpendale (Eds.), *Data-driven storytelling* (pp. 107–133). CRC Press. <https://doi.org/10.1201/9781315281575-5>
- Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17(2), 191–204. <https://doi.org/10.2307/1165169>
- Behrens, J. T., DiCerbo, K., Murphy, D., & Robinson, D. (2013, April 28–30). *Conceptual frameworks for reporting results of assessment activities* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.
- Bennett, E., & Folley, S. (2019). Four design principles for learner dashboards that support student agency and empowerment. *Journal of Applied Research in Higher Education*, 12(1), 15–26. <https://doi.org/10.1108/JARHE-11-2018-0251>
- Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6), 706–732. <https://doi.org/10.3102/1076998618784700>
- Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 41–50). Association for Computing Machinery.
- Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Praeger.
- Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of Project asTTle outputs design* (asTTle Tech. Rep. #15). University of Auckland. https://www.researchgate.net/publication/237549707_Reporting_Assessment_Information_to_Teachers_Report_of_Project_asTTle_Outputs_Design_Technical_Report_15
- Brown, G. T. L., Kannan, P., Sinharay, S., Zapata-Rivera, D., & Zenisky, A. L. (2023). Challenges and opportunities in score reporting: A panel of personal perspectives. *Frontiers in Education*, 8, 1211580. <https://doi.org/10.3389/feduc.2023.121158>
- Brown, G. T. L., O’Leary, T. M., & Hattie, J. A. C. (2019). Effective reporting for formative assessment: The asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 107–125). Routledge.
- Bulut, O., Cutumisu, M., Singh, D., & Aquilina, A. M. (2020). Guidelines for generating effective feedback from e-assessments. *Hacettepe University Journal of Education*, 35, 60–72. <https://doi.org/10.16986/HUJE.2020063705>
- Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008, December). *Web content accessibility guidelines (WCAG) 2.0*. <https://www.w3.org/TR/WCAG20/>
- Clauser, B. E. (2019, April 4–8). *A history of classical test theory* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, Canada.

- Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Praeger.
- Cope, B., & Kalantzis, M. (2015.) Sources of evidence of learning: Learning and assessment in the era of big data. *Open Review of Educational Research*, 2(1), 194–217. <https://doi.org/10.1080/23265507.2015.1074869>
- Corrin, L., Kennedy, G., French, S., Buckingham Shum S., Kitto, K., Pardo, A., West, D., Mirriahi, N., & Colvin, C. (2019). *The ethics of learning analytics in Australian higher education. A discussion paper*. University of Melbourne CSHE. https://melbourne-cshe.unimelb.edu.au/__data/assets/pdf_file/0004/3035047/LA_Ethics_Discussion_Paper.pdf
- Daniels, L. M., & Bulut, O. (2019). Students' perceived usefulness of computerized percentage-only vs. descriptive score reports: Associations with motivation and grades. *Journal of Computer Assisted Learning*, 36(2), 199–208. <https://doi:10.1111/jcal.12398>
- De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology*, 10, Article 1280. <https://doi.org/10.3389/fpsyg.2019.01280>
- Du, X., Yang, J., Shelton, B. E., Hung, J.-L., & Zhang, M. (2021) A systematic meta-review and analysis of learning analytics research. *Behaviour & Information Technology*, 40(1), 49–62. <https://doi.org/10.1080/0144929X.2019.1669712>
- Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Routledge.
- Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 603–623). Routledge.
- Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media.
- Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019.) Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy*, 34(4), 470–499. <https://doi.org/10.1080/02680939.2018.1460493>
- Forte Fast, E., & Tucker, C. (2001, April 10–14). *Redesign of the student assessment reporting system in Connecticut* [Paper presentation]. American Educational Research Association Annual Meeting, Seattle, WA, United States.
- Gerst, M. D., Kenney, M. A., Baer, A. E., Speciale, A., Wolfinger, J. F., Gottschalck, J., Handel, S., Rosencrans, M., & Dewitt, D. (2019). Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks. *Weather, Climate, and Society*, 12(1), 117–133. <https://doi.org/10.1175/WCAS-D-18-0094.1>
- Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17(2), 145–220. https://doi.org/10.1207/s15324818ame1702_3

- Gotch, C. M., & Roduta Roberts, M. (2018). A review of recent research on individual-level score reports. *Educational Measurement: Issues and Practice*, 37, 46–54. <https://doi.org/10.1111/emip.12198>
- Hambleton, R. K. (1998). Enhancing the validity of NAEP achievement level score reporting. In M. L. Bourque (Ed.), *Proceedings of the Achievement Levels Workshop* (pp. 77–98). National Assessment Governing Board.
- Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scale and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192–205). Allyn & Bacon.
- Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990–1998. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 133–155). National Assessment Governing Board.
- Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). National Center for Research on Evaluation, Standards, and Student Teaching.
- Hambleton, R. K., & Zenisky, A. L. (2003). Advances in criterion-referenced testing methods and practices. In C. R. Reynolds & R. W. Kamphaus (Eds.), *The handbook of psychological and educational assessment* (2nd ed., pp. 377–404). The Guilford Press.
- Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology. Vol. 3: Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. <https://doi.org/10.1037/14049-023>
- Hambleton, R. K., & Zenisky, A. L. (2017). Score reporting and interpretation. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (2nd ed., pp. 127–142). Chapman and Hall/CRC.
- Hamilton, L. S., Stecher, B. M., & Yuan, K. (2009.) *Standards-based reform in the United States: History, research, and future directions*. Center on Education Policy.
- Harold, J., Lorenzoni, I., Shipley, T., & Coventry, K. (2016). Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change*, 6, 1080–1089. <https://doi.org/10.1038/nclimate3162>
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hattie, J. (2010). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*, 1(5), 1–15. <https://oerj.webspace.durham.ac.uk/wp-content/uploads/sites/249/2021/07/4.pdf>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hegarty, M. (2011). The cognitive science of visual–spatial displays: Implications for design. *Topics in Cognitive Science*, 3, 446–474. <https://doi.org/10.1111/j.1756-8765.2011.01150.x>

- Hegarty, M. (2019). Advances in cognitive science and information visualization. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 19–34). Routledge.
- Hegarty, M., Friedman, A., Boone, A., & Barrett, T. J. (2016). Where are you? The effect of uncertainty and its visual representation on location judgments in GPS-like displays. *Journal of Experimental Psychology: Applied*, 22(4), 381–392. <https://doi.org/10.1037/xap0000103>
- Hooper, L. (2021). *How to tell a story with data: A guide for beginners*. Venngage. <https://venngage.com/blog/data-storytelling/>
- Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011, November). *Research on graph comprehension and data interpretation: Implications for score reporting* [Paper presentation]. ETS Score Reporting Conference, Princeton, NJ, United States.
- International Test Commission. (2013). *International guidelines for test use* (version 1.2). Retrieved March 2, 2024, from https://www.intestcom.org/files/guideline_test_use.pdf
- Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003-11). U.S. Department of Education, Institute of Education Sciences.
- Jaeger, R.M., Gorney, B., Johnson, R. L., Putnam, S. E., & Williamson, G. (1993). *Designing and developing effective school report cards: A research synthesis*. Center for Research on Education Accountability and Teacher Evaluation, Western Michigan University.
- Johannsen, I. M., Lassonde, K. A., Wilkerson, F., & Schaab, G. (2018). Communicating climate change: Reinforcing comprehension and personal ties to climate change through maps. *The Cartographic Journal*, 55(1), 85–100. <https://doi.org/10.1080/00087041.2017.1386834>
- Jonsson, A. (2012). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, 14(1), 63–76. <https://doi.org/10.1177/1469787412467125>
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.
- Kane, M. T. (2015). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). Routledge.
- Kannan, P., & Zapata-Rivera, D. (2022). Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: Informing the design and development of learning analytic dashboards. *Frontiers in Education*, 7, 913594. <https://doi.org/10.3389/feduc.2022.913594>
- Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educational Assessment*, 23(3), 173–194. <https://doi.org/10.1080/10627197.2018.1477584>
- Katz, I. R. (2019). Foreword. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. xiii–xiv). Routledge.
- Kay, J. (1997). Learner know thyself: Student models to give learner control and responsibility. In Z. Halim, T. Ottomann, & Z. Razak (Eds.), *Proceedings of*

- International Conference on Computers in Education* (pp. 17–24). AACE. <https://www1.inhatc.ac.kr/seihoon/papers/learning/Kay97.pdf>
- Khasnabish, S., Burns, Z., Couch, M., Mullin, M., Newmark, R., & Dykes, P. C. (2020). Best practices for data visualization: Creating and evaluating a report for an evidence-based fall prevention program. *Journal of the American Medical Informatics Association*, 27(2), 308–314. <https://doi.org/10.1093/jamia/ocz190>
- Kirylo, J. D. (2018). The opt-out movement and the power of parents. *Phi Delta Kappan*, 99(8), 36–40. <https://doi.org/10.1177/0031721718775676>
- Knupp, T., & Ansley, T. (2008, March 25–27). *Online, state-specific assessment score reports and interpretive guides* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.
- Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311846.001.0001>
- Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.
- Levine, R., Rathbun, A., Selden, R., & Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress constituents survey and focus groups* (NCES 98–521). U.S. Department of Education, Office of Educational Research and Improvement.
- Linares-Vásquez, M., Moran, K., & Poshyvanyk, D. (2017, September). Continuous, evolutionary and large-scale: A new perspective for automated mobile app testing. In *2017 IEEE International Conference on Software Maintenance and Evolution* (pp. 399–410). IEEE.
- Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.
- Linn, R. L. (Ed.). (1989). *The American Council on Education/Macmillan series on higher education. Educational measurement* (3rd ed.). American Council on Education and Macmillan.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, 31(6), 3–16. <https://doi.org/10.3102/0013189X031006003>
- Luecht, R. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, 14, 1–38. <https://www.testpublishers.org/jatt-volume-14>
- MacIver, R., Anderson, N., Costa, A. C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, 22(2), 149–164. <https://doi.org/10.1111/ijsa.12065>
- Measurement Research Associates, Inc. (2010). *Using the very useful Wright map*. Retrieved August 26, 2015, from <http://www.rasch.org/mra/mra-01-10.htm>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.
- Milligan, S. K., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. *Journal of Learning Analytics*, 3(2), 88–115.

- Mislevy, R. J. (1998). Implications of market-basket-reporting for achievement level setting. *Applied Measurement in Education*, 11(1), 49–63. https://doi.org/10.1207/s15324818ame1101_3
- Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, 4(1), 11–48. <https://doi.org/10.5281/zenodo.3554641>
- Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20. <https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Mosier, C. I. (1951). Batteries and profiles. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 764–808). American Council on Education.
- National Center for Education Statistics. (2013). *The Nation's Report Card: Trends in academic progress 2012* (NCES 2013–456). Institute of Education Sciences, U.S. Department of Education.
- National Commission for Certifying Agencies. (2014). *Standards for the accreditation of certification programs*. Institute for Credentialing Excellence.
- National Commission on Excellence in Education. (1983). *A Nation at Risk: The imperative for educational reform*. U.S. Government Printing Office.
- National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting*. National Academies Press.
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 101, Stat. 1425 (2002).
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- O'Donnell, F. (2020). *What's in a label? Unpacking the meaning of achievement labels from tests* [Doctoral dissertation, University of Massachusetts Amherst]. 1856. <https://doi.org/10.7275/15558737>
- O'Donnell, F., Ong, T. Q., & Feinberg, R. (2021, May 18–June 11). Advances in reporting results on medical education assessments. In C. Runyon & S. Somay (Chairs), *Advancing assessment in medical education* [Coordinated session]. National Council on Measurement in Education Annual Meeting.
- O'Donnell, F., & Sireci, S. G. (2019). Score reporting issues for licensure, certification, and admissions programs. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 77–90). Routledge.
- O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment*, 27(1), 1–26. <https://doi.org/10.1080/10627197.2021.2016388>
- O'Donnell, F., & Zenisky, A. L. (2020). Digital module 21: Results reporting for large-scale assessments. *Educational Measurement: Issues and Practice*, 39, 137–138. <https://doi.org/10.1111/emip.12408>
- O'Leary, T. M., Hattie, J. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues & Practice*, 36(2), 16–23. <https://doi.org/10.1111/emip.12141>

- Oliva-Córdova, L. M., Garcia-Cabot, A., & Amado-Salvatierra, H. R. (2021). Learning analytics to support teaching skills: A systematic literature review. *IEEE Access*, 9, 58351–58363. <https://doi.org/10.1109/ACCESS.2021.3070294>
- Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, 3(1), 29–54. <https://doi.org/10.1186/s41235-018-0120-9>
- Park, Y., & Jo, I. (2019). Factors that affect the success of learning analytics dashboards. *Educational Technology Research and Development*, 67(6), 1547–1571. <https://doi.org/10.1007/s11423-019-09693-0>
- Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, 9(1), 17–17. <https://doi.org/10.1186/s40536-020-00092-z>
- Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, 14(1), 36–49. <https://doi.org/10.1037/1076-898X.14.1.36>
- Reese, T. J., Segall, N., Del Fiore, G., Tonna, J. E., Kawamoto, K., Weir, C., & Wright, M. C. (2020). Iterative heuristic design of temporal graphic displays with clinical domain experts. *Journal of Clinical Monitoring and Computing*, 35(5), 1119–1131. <https://doi.org/10.1007/s10877-020-00571-2>
- Roberts, M., & Gierl, M. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, 29(3), 25–38. <https://doi.org/10.1111/j.1745-3992.2010.00181.x>
- Roduta Roberts, M., Gotch C. M., & Lester, J. N. (2018). Examining score report language in accountability testing. *Frontiers in Education*, 3(42), 1–17. <https://doi.org/10.3389/educ.2018.00042>
- RStudio. (2020). *Shiny: Easy web applications in R*. <https://shiny.rstudio.com/>
- Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Lawrence Erlbaum Associates.
- Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2019). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, 29(1), 682–692.
- Schwendimann, B., Rodriguez-Triana, M., Vozniuk, A., Prieto, L., Boroujeni, M., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, 10(1), 30–41. <https://doi.org/10.1109/tlt.2016.2599522>
- Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). ACM.
- Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Chapman & Hall/CRC. <https://plotly-r.com>

- Simmons, C., & Mwalimu, M. (2000). What NAEP's publics have to say. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 184–219). National Assessment Governing Board.
- Sinharay, S., Haberman, S. A., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21–28.
- Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2019). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 35–49). Routledge.
- Slater, S., Livingston, S. A., & Silver, M. (2019). Score reports for large-scale testing programs: Managing the design process. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 91–106). Routledge.
- Tannenbaum, R. J. (2019). Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 9–18). Routledge.
- Thorndike, R. (Ed.). (1971). *Educational measurement* (2nd ed.). American Council on Education.
- Thudt, A., Gschwandtner, T., Walny, J., Dykes, J. & Stasko, J. (2018). Exploration and explanation in data-driven storytelling. In N. Riche, C. Hurter, N. Diakopoulos, & S. Carpendale (Eds.), *Data-driven storytelling* (pp. 59–83). CRC Press. <https://doi.org/10.1201/9781315281575-5>
- Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press.
- Tufte, E. R. (1990). *Envisioning information*. Graphics Press.
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2015). *NAEP science—map of selected item descriptions on the NAEP Science Scale—Grade 8*. Retrieved August 25, 2015, from <http://nces.ed.gov/nationsreportcard/itemmaps/index.asp>
- Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation*, 43, 24–39. <https://doi.org/10.1016/j.stueduc.2014.04.004>
- Verbert, K., Govaerts, S., Duval, E., Santos, J., Van Assche, F., Parra, G., & Klerkx, J. (2013). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing*, 18(6), 1499–1514. <https://doi.org/10.1007/s00779-013-0751-2>
- Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, 21(1), 14–23. <https://doi.org/10.3102/0013189X021001014>
- Wainer, H. (1997a). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Copernicus Books.
- Wainer, H. (1997b). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics*, 22(1), 1–30. <https://doi.org/10.3102/10769986022001001>

- Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton University Press.
- Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate, and control uncertainty through graphical display*. Princeton University Press.
- Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement*, 36(4), 301–335. <https://doi.org/10.1111/j.1745-3984.1999.tb00559.x>
- Waters, B. K. (1997). Army Alpha to CAT-ASVAB: Fourscore years of military personnel selection and classification testing. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 187–203). Greenwood Press.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. <https://ggplot2.tidyverse.org>
- Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). Statistics and computing. Springer.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 1–15. <https://doi.org/10.3389/fpsyg.2019.03087>
- Xiao, Y., Veldkamp, B., & Liu, H. (2022). Combining process information and item response modeling to estimate problem-solving ability. *Educational Measurement: Issues and Practice*, 41, 36–54. <https://doi.org/10.1111/emip.12474>
- Yen, W. M. (1987). *A Bayesian/IRT index of objective performance* [Paper presentation]. Meeting of the Psychometric Society, Montreal, Canada.
- Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS Score Reporting Conference* (ETS Research Report No. RR-11-45). ETS. <https://www.ets.org/Media/Research/pdf/RR-11-45.pdf>
- Zapata-Rivera, D. (2021). Open student modeling research and its connections to educational assessment. *International Journal of Artificial Intelligence in Education*, 31, 380–396. <https://doi.org/10.1007/s40593-020-00206-2>
- Zapata-Rivera, D., Andrews-Todd, J., & Oliveri, M. E. (2021). Communicating assessment information in the context of a workplace formative task. *The Journal of Writing Analytics*, 5, 324–341. <https://doi.org/10.37514/JWA-J.2021.5.1.10>
- Zapata-Rivera, D., Kannan, P., & Zwick, R. (2019). Communicating measurement error information to teachers and parents. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 63–73). Routledge.
- Zapata-Rivera, D., & Katz, R. I. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice*, 21(3), 442–463.
- Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (ETS Research Memorandum RM-12-20). ETS.
- Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results.

- Educational Assessment*, 21(3), 215–229. <https://doi.org/10.1080/10627197.2016.1202110>
- Zenisky, A. L. (2015). Visual displays for reporting test data: Making sense of test performance. In M. McCrudden, G. Schraw, & C. Buckendahl (Eds.), *Use of visual displays in research and testing: Coding, interpreting, and reporting data* (pp. 299–332). Information Age Publishing.
- Zenisky, A. L., Delton, J., & Hambleton, R. K. (2006). *State reading content specialists and NAEP data displays* (Report No. 598). University of Massachusetts, Center for Educational Assessment.
- Zenisky, A. L., & Hambleton, R. K. (2007). *Navigating ‘The Nation’s Report Card’ on the World Wide Web: Site user behavior and impressions* [Technical report]. Comprehensive Evaluation of NAEP. [Also Center for Educational Assessment Report No. 625]
- Zenisky, A. L., & Hambleton, R. K. (2012a). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, 31(2), 21–26. <https://doi.org/10.1111/j.1745-3992.2012.00231.x>
- Zenisky, A. L., & Hambleton, R. K. (2012b). From “here’s the story” to “you’re in charge”: Developing and maintaining large-scale online test and score reporting resources. In M. Simon, M. Rousseau, & K. Ercikan (Eds.), *Improving large-scale assessment in education* (pp. 175–185). Routledge.
- Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 585–602). Routledge.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education*, 22(4), 359–375. <https://doi.org/10.1080/08957340903221667>
- Zenisky, A. L., Hambleton, R. K., & Smith, Z. R. (2006). *Do math educators understand NAEP score reports? Evaluating the utility of selected NAEP data displays* (Report No. 587). University of Massachusetts, Center for Educational Assessment.
- Zenisky, A. L., Keller, L. A., & Park, Y. (2019). Reporting student growth: Challenges and opportunities. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 50–62). Routledge.
- Zenisky, A. L., Mazzeo, J., & Pitoniak, M. J. (2016). Towards improving the reporting of test score results: Lessons learned from the evolution of NAEP reports. In C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 335–355). Guilford.
- Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O’Donnell, F., Wells, C. S., Padellaro, F., Jung, H. J., Banda, E., Pham, D., Hong, S. E., Park, Y., Botha, S., Lee, M., & Garcia, A. (2018). *Massachusetts adult proficiency tests for college and career readiness: Technical manual* (Center for Educational Assessment Research Report No. 974). Center for Educational Assessment.

- Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, 20, 79–87. <https://doi.org/10.1016/j.pse.2014.11.003>
- Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, 19(2), 116–138. <https://doi.org/10.1080/10627197.2014.903653>.

NOTE

1. Throughout this chapter, we honor the legacy of our colleague and this chapter's third author, Ronald K. Hambleton, by incorporating specific quotations drawn from his extensive body of work on this topic. Professor Hambleton's countless contributions have significantly shaped our field and continue to frame our discourse. His passing on April 28, 2022, was a profound loss, but his voice and influence endure through his writings, which we are privileged to share here.

Realizing Fairness Through Accessibility for All Test Takers and for Specific Groups

Michael C. Rodriguez

University of Minnesota Twin Cities

Martha L. Thurlow

University of Minnesota Twin Cities (Retired)

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), hereafter referred to as the *Standards*, are forward thinking in projecting fairness as a foundational concept for all educational and psychological measurement. The *Standards* described several general views of fairness, including equitable treatment during the testing process, absence of measurement bias, validity of individual test score interpretation and use, and access to the intended measured constructs and achievement domains. Increasingly, the concept of *accessibility* has been identified as a key aspect of realizing fairness in educational and psychological testing. Overall, a fair test maximizes construct and domain-relevant information, regardless of irrelevant individual characteristics and test contexts, to ensure valid test score interpretation and use for all test takers.

Zwick (this volume) notes the interconnectedness of the three foundational chapters in the *Standards*—validity, reliability, and fairness—as well as their interrelationships with concepts such as bias, construct-irrelevant variance, opportunity to learn, standardization, universal design for assessment (UDA), and others. In the foundational discussion of fairness in the *Standards* is the recognition that fairness is a broad concept that includes an overarching standard about the entire testing process being designed to minimize construct-irrelevant variance to promote valid score interpretations. The fairness standards are organized in four clusters that address (a) test design, development, administration, and scoring; (b) validity of score interpretations; (c) accommodations to remove construct-irrelevant barriers; and (d) safeguards against inappropriate score interpretations. Zwick moves beyond these concepts to consider fairness in the context of opportunity to learn and selection (admissions to higher education and employment). She also argues that fairness must be addressed through a larger validity argument which provides “a chain of reasoning using documented evidence to support the fairness of the intended uses and interpretations of test scores” (Zieky, 2016, as cited in Zwick, this volume, p. ##).

For the first time, the *Standards* eliminated chapters devoted to individuals with specific characteristics, such as those with disabilities and those from different cultural and linguistic backgrounds. Instead, a focus on all test takers was woven throughout the chapters with an emphasis on meeting the needs of all test takers throughout test design, development, implementation, reporting, interpretation, and use. Realizing accessibility is critical in this approach, which is much more than just providing accommodations to some students. Accessibility is a concept that has emerged slowly and been applied to educational and psychological tests as it has become clear that test takers include culturally and linguistically diverse groups of individuals.

The purpose of this chapter is to apply the theory and concepts of a fairness argument to all test-taking populations, both in general and for specific groups of test takers. We explore these through the lens of accessibility and what it means in general for all test takers, as well as through the perspective of specific groups, including individuals with disabilities, English learners, those from different racialized and ethnicized groups, and those living in poverty. We address fairness for all types of assessments, including tests of achievement most often used in preschool through Grade 12 (P–12) school settings

(e.g., used for formative, interim, or summative purposes), higher education classroom settings, campus-wide assessments of student learning outcomes, admissions tests (e.g., ACT, SAT, GRE, Law School Admission Test [LSAT]), and licensure and certification exams. At the end of most sections, we provide implications for practice, or practice and research, and summarize some of the key guidance promoted and informed by the scholars reviewed in those sections.

Accessibility is the condition where test takers have “unobstructed opportunity to demonstrate their standing on the construct(s) being measured” (AERA et al., 2014, p. 49). A related approach, which has become a common component of test design principles, is UDA, borrowed from principles of architecture regarding “test design that seeks to maximize accessibility for all intended examinees” (AERA et al., 2014, p. 50), to be responsive to test-taker characteristics and the testing context(s). In addition, we expand the notions of UDA by including the principles of universal design for learning (UDL) created by CAST (<https://www.cast.org/>), a broader conceptualization of access that acknowledges that access to content and assessment tasks is necessary, but not sufficient; test takers need access to learning opportunities. Together, a fair test “reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population” (AERA et al., 2014, p. 50), regardless of racialized status, ethnicity, gender, age, socioeconomic status, special education status, or cultural or linguistic background. We use the term racialized status, rather than race, acknowledging that race is a social construct and that minoritized individuals are ascribed to racial groups, marginalized, and subjected to racism and oppression (in many ways, this includes ethnicity). As employed by Russell (2024), the term racialized denotes “the active process of creating categories into which humans are membered” (p. viii).

Furthermore, fairness is of concern in each stage of test design, development, administration, scoring, interpretation, and use. The fairness chapter in the *Standards* addresses four general views, including (a) fair and equitable treatment of test takers, (b) fairness as access to the measured constructs, (c) fairness as lack of measurement bias, and (d) fairness as validity of score interpretation and use; the latter two views are addressed comprehensively by Zwick (this volume). Finally, we explore the promise of culturally and linguistically responsive assessment to provide accessibility and achieve greater test fairness.

Fairness is often discussed in terms of test-taker characteristics needing additional or unique considerations in test development (see Huff et al., this volume), administration and scoring (see Shermis et al., this volume), and reporting (see Zenisky et al., this volume). However, it is important to acknowledge that fairness challenges are less about the test-taker characteristics and more about the quality, inclusiveness, and responsiveness of test design and implementation, where challenges result when test characteristics interact with test-taker characteristics in unintended (or unanticipated) ways (Kettler et al., 2018), thus the focus on accessibility and universal design principles. We agree that radical goals can be achieved “when educational tests focus on promoting the success of *all* students” (Sireci, 2021, p. 7).

FAIRNESS IN EDUCATIONAL TESTS AND ASSESSMENTS

For our purpose, we focus on fairness in the measurement of education achievement and do not directly address more purely psychological assessments, employment tests such as cognitive ability or personality tests, assessments of job performance, or other tests typically in the purview of clinical or industrial-organizational psychology. Both educational and psychological tests and assessments have fairness concerns (e.g., see Geisinger, 2015), and although many of the issues and contexts have parallels in psychological testing, our focus is more directly applied to testing in education settings or for education purposes.

Fairness and accessibility are concerns for all types of educational tests and assessments and all proposed purposes (including those for which validity evidence may or may not exist). For tests with more high-stakes outcomes, more rigorous evidence is needed to support fairness and accessibility, as with validity evidence. This includes tests of achievement for interim and summative purposes with implications for individuals (see Brookhart & DePascale, and Ho & Polikoff, both this volume); tests used for admissions to higher education, such as the ACT, SAT, GRE, LSAT, Medical College Admission Test [MCAT], GMAT, and others (see Camara et al., this volume); tests used for licensure and certification (see Margolis et al., this volume); and tests used for assessment of so-called social and emotional, character, behavioral, or intra- and interpersonal skills (see Kyllonen & Zu, this volume).

Our purpose is to apply the theory and concepts of fairness (see Zwick, this volume) to test-taking populations, both in general (i.e., accessibility for all test takers) and for specific groups (e.g., English learners, individuals with disabilities). “Regardless of the purpose of testing, the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure” (AERA et al., 2014, p. 51). Of course, such opportunity to demonstrate standing is provided by the test developer and administrator; the test should enable such access given the needs of the test taker.

A comment is needed at this point regarding terms used throughout the chapter, including test and assessment and construct and domain. The authors of the *Standards* described a test as a tool or procedure to systematically sample test-taker behavior in a specified domain in a standardized process. Assessment includes a broader set of systematic methods for obtaining information about a person’s characteristics or performance; assessment includes processes to collect relevant evidence, interpret the resulting information, and inform a decision (Russell, 2022). Both include systematic processes and both require domain specifications to support intended interpretations and uses.

In this chapter, we use the word *test* when we mean a tool containing items and tasks administered and scored in systematic ways to measure (a quantitative activity) what a person knows and can do relative to the targeted domain; we use *assessment* when we mean the broader range of processes, both qualitative and quantitative, informal and

formal, to similarly collect evidence of what a person knows and can do. In addition, a *construct* is a specific “concept or characteristic that a test is designed to measure” (AERA et al., 2014, p. 217), whereas a content *domain* is “the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test, represented in detailed test specifications and often organized into categories by which items are classified” (AERA et al., 2014, p. 218). Although construct and domain appear to be used interchangeably in most educational measurement texts, we prefer the term domain in most education contexts because the tests we write about primarily include measures of achievement, with systematic reference (inference) to what people know and can do as reflected in test specifications (often including many constructs). We use the terms employed by other authors when representing their specific arguments and claims.

Fairness and the Choice to Test

The fact that a test is used presupposes the use of an inclusive decision process that determined that testing is an appropriate, meaningful, and useful method to collect evidence to support decision-making. At this early stage in the testing process, fairness considerations abound. Many decisions occur on a regular basis employing information about knowledge and skills across the life span, from early childhood through careers (e.g., progression through P–12 grade levels, placement in special education or gifted/talented programs, admissions to higher education institutions, or licensure and certification). Evidence can be collected through a wide range of informal to formal qualitative and quantitative approaches. Especially in accountability-based settings, decisions are improved when they are informed; yet what qualifies as evidence is value laden and culture bound. Drawing heavily on the work of Messick (1975, 1980, 1981), Kane (1992) argued that “the legitimacy of test use rests on the assumptions about the possible outcomes (intended and unintended) of the decision to be made and on the values associated with these different outcomes” (p. 530). Messick (1965) argued for attention to regulating ethical standards to justify test use.

Values enter our considerations in a number of ways, including the framing of the decisions that need to be made in education settings, as well as the inclusivity of who is at the table to determine the relevance of impending decisions. When testing is introduced as a means through which evidence can be collected, additional questions must be addressed. Those questions include what sources of evidence to gather to inform decisions; what we value as evidence can vary substantially and impact the determination of what to measure and how to measure it. Such questions are addressed in the final section, where we discuss accessibility through cultural and linguistic responsiveness.

Because fairness addresses the treatment of individuals and groups, a decision to test must consider the appropriateness of test use for all individuals and groups in the test-taker population, including those who may be tested in the near future. This includes opportunity to learn, prior test experience, access to test preparation, access to the test content itself in terms of cultural and linguistic accessibility and language of testing, and ability to respond to items or tasks as required by the test administration procedures.

“A proposal to use a test for a particular purpose, then, should be appraised in light of the probable future consequences of the testing, not only in terms of what it entails but also in terms of what it makes more likely” (Messick, 1975, p. 962). Zwick (this volume) discusses the possibility that, in some cases, the decision to not use a test is most appropriate. For example, if the use of a diagnostic reading test for the purpose of identifying students who need reading remediation identifies 80% to 90% of the students of color in a large population of mostly nonimmigrant children but only 10% of White students, a decision to not employ such a test is warranted, regardless of technical or psychometric quality.

Implications for Practice

Practitioners should secure diverse participation and include key stakeholders in the decision-making process regarding whether to use a test to inform certain decisions. Fairness requires consideration of whether testing is appropriate for all potential test takers and whether a test is of high enough technical or psychometric quality across the test-taker population to warrant use. Test developers and test users should collaboratively contend with the request of Messick (1975) to appraise the probable uses and future consequences of the testing program.

FAIRNESS AS ACCESSIBILITY FOR ALL TEST TAKERS

Accessibility Defined and Explored

“Optimal accessibility is implicitly promised to all” test takers (Kettler et al., 2018, p. 1), as is fairness in testing. Accessibility principles apply to instruction as the opportunity for individuals to learn the disciplinary knowledge and practices through an intended curriculum and to assessment as the opportunity for test takers to demonstrate the disciplinary knowledge and practices required by the test content (Kettler et al., 2009). Full access to instruction and assessment minimizes bias and increases fairness (Kettler et al., 2018), connecting the instruction and assessment accessibility principles directly to the notion of accessibility in the *Standards*.

Equity in access has been an important objective of federal education legislation (e.g., the Civil Rights Act of 1964 and Elementary and Secondary Education Act of 1965 [ESEA]) and case law (e.g., *Brown v. Board of Education*, 1954), particularly in the context of special education (e.g., Education for All Handicapped Children Act of 1975, reauthorized as the Individuals With Disabilities Education Act [IDEA] in 1990). A large body of legislative and federal laws and administrative rules protects inclusion and equitable access to instruction and assessments.

Prior to the 2014 *Standards*, researchers and practitioners were exploring and clarifying the meaning of accessibility. Willingham and Cole (1997) indicated that fair tests measure the same construct for all test takers with the same meaning for all individuals in the population. More specifically, test accessibility is “the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge

of the target construct” (Beddow, 2012, p. 98). These definitions are consistent with the notions of fairness in the *Standards*, and here we embrace the definition offered by Beddow.

Test score error variance due to accessibility barriers could originate from physical, perceptive, receptive, emotive, and cognitive sources, sources that may be linked to test or item characteristics, administrator characteristics, or administration contexts and practices (Beddow, 2012; Elliott et al., 2018). In its most general sense, “accessibility—defined as the extent to which a product, environment, or system eliminates barriers and permits equal use of components and for a diverse population of individuals—is necessary for effective instruction and fair testing” (Elliott et al., 2018, p. 1). In the broader context of fairness, we include potential accessibility-limiting sources due to social, cultural, and linguistic contexts and other construct-irrelevant characteristics of test takers, not directly addressed in the sources previously discussed. In particular, accessibility principles that recognize the unique challenges of ensuring that assessments are accessible for Indigenous populations are promising. For example, Hawaiian educators developed principles for Hawaiian culture-based education focused on the purpose of the test, the design and content of the test, the context in which the test is conducted, and how the results are used (Nā Lau Lama, 2007). This work resulted in tests that embodied shared meanings, aligned with the Indigenous elements in the curriculum and language of instruction (in some cases, bilingual instruction), and sustained culturally grounded practices such as *hō’ike*, which includes meaningful authentic tasks (Nā Lau Lama, 2007).

Accessibility Factors That Affect Fairness

There are three components to accessibility relevant to test fairness, including access skills, target skills, and adaptations (Ketterlin-Geller, 2008; Kettler et al., 2018). Test takers need access skills to allow them to perform well on tests, skills that we often take for granted and assume are present prior to testing. In most cases, such skills are developed through test-taking experiences, including informal classroom tests, and sometimes through test preparation courses or training programs. In some cases, test preparation can help level the playing field and secure an equitable level of test-taking skills across test takers. A common recommendation when employing new item types or preparing test takers with little test-taking experience is to provide opportunities for students to practice so that they can know what is expected and how to interact and respond to test items. However, there are limits to test preparation, where some activities may cross ethical boundaries and cause inferences to be invalidated and, as such, not support fairness as accessibility. These include, as the most egregious, releasing and practicing on operational items or focusing effort on studying released test content and items over a deeper exploration of the learning objectives emerging from content standards on which the test is based. However, as suggested previously, some practice may be useful for those unfamiliar with the item formats or tasks.

Test takers need to explore, understand, and employ disciplinary knowledge and skills required to perform well on a specific test given the targeted knowledge and skills intended to be measured. Much of this chapter addresses accessibility to target knowledge and skills. There is a substantial literature on the accessibility of instruction to support opportunity to learn (Kettler et al., 2018). And in some cases, accommodations or modifications are needed to facilitate access to both instruction and testing. One area of persistent challenge is the degree to which the disciplinary knowledge and practices are culturally or linguistically relevant to the test taker—making explicit the limited degree to which a construct or domain is defined and represented in instruction, in practice, and on the test.

What do we mean by “the degree to which a test and its constituent item set permit the test taker to demonstrate his or her knowledge of the target construct” (Beddow, 2012, p. 98)? Our operational definition of accessibility has a hidden component of significant ambiguity—the target construct. We may simply accept the test developer’s target domain (construct) as declared in the test specifications. Or, we might ask: Does the domain definition prohibit the possibility of allowing some test takers to demonstrate their disciplinary knowledge and skills from culturally and linguistically relevant disciplinary knowledge and practices? In the United States, much of the curriculum employed in P–12 and higher education is deeply entrenched in middle-class Euro-centered and colonialist history and perspectives of what people should know and be able to do. Consider the practice of wild rice cultivation (a centuries-long tradition in American Indian communities in Minnesota), which embodies disciplinary knowledge and practices from science, mathematics, history, and agriculture and provides rich opportunities to explore reading, writing, and storytelling; test items that place subject matter disciplinary knowledge in contexts known to students make them relevant and accessible (see, for example, Randall, 2021). Culturally and linguistically relevant disciplinary knowledge and practices are important sources to secure deeper accessibility and fairness in testing via fair domain definitions. Construct and domain definitions play a critical role in test fairness and validation (Randall et al., 2023).

Others also have argued that fairness is enhanced through principles of access via adaptations, including test modifications or accommodations (e.g., Rose et al., 2005; Thompson et al., 2004; Thurlow et al., 2016; Thurlow, Liu, et al., 2013; Zieky, 2016), whereas, in some cases, fairness and validity are reduced in the presence of barriers to access or sources of construct-irrelevant variance. We emphasize the applicability of accessibility principles across the life span, in the many settings and purposes of testing. In those contexts, there are multiple levels at which tests are administered, interpreted, and used, including interpretation and use at the individual (e.g., GRE) and group (e.g., National Assessment of Educational Progress [NAEP]) levels. An important component of accessibility, when no other modifications are possible, is the appropriate provision and use of accommodations.

There are many aspects of test design, administration, and reporting and use that potentially interact with test-taker characteristics that may call into question the fairness

of test accessibility. Although test developers work to attend to every possible personal characteristic or experience in the testing process, some groups have historically faced discrimination and significant barriers to education and career opportunities. In a similar way, there are as many testing conditions and contexts under which fairness can be realized as there are potentials for barriers and interference for fair test interpretation and use. Here, we briefly describe several of the more prominent and potentially promising conditions under which fairness can be enhanced and supported. We also highlight the importance of accessibility features and accommodations for supporting fair testing and note the role of testing consequences.

Implications for Practice

Test developers and users should ensure that all test takers have adequate test-taking access skills and have access to appropriate test preparation that is available freely or at low cost. Test developers must make explicit the source of target domains and explore alternative sources of disciplinary knowledge and practices, vis-à-vis the diversity within the target test-taker population. As necessary, when additional adaptations are not possible, accommodations may be identified that fulfill the goals of access and fairness.

FAIRNESS AS ACCESSIBILITY FOR SPECIFIC GROUPS OF TEST TAKERS

Much of the history of thinking about accessibility as an avenue to fairness has grown out of concerns about fairness of testing results for specific groups of individuals. Considerations of fairness have been enhanced through principles of access (e.g., Nā Lau Lama, 2007; Rose et al., 2005; Thompson et al., 2004; Thurlow et al., 2016; Thurlow, Liu, et al., 2013; Zieky, 2016) that identify ways to reduce the presence of barriers to access or sources of construct-irrelevant variance. Accessibility principles can be applied across the life span, in the many settings and purposes of testing. In those contexts, there are multiple levels at which tests are administered, interpreted, and used, at both the individual and the group levels. An important component of accessibility has traditionally focused on the provision of accommodations and their appropriate use.

Initially, the push for thinking about how to make tests available to individuals grew out of the need to identify ways to allow individuals who were blind, who were deaf, or who had certain types of learning disabilities to take tests that would produce results that could be interpreted in the same way as the results of other individuals. We provide here an overview of the progression of work starting in the 1980s. It was not until the 1990s that discussions about accessibility began for English learners. Again, these efforts were focused almost always on accommodations. In the early 2000s, a number of states were engaged in enhanced assessment grants to transform accommodation practices and support item development and test delivery; these efforts were brought to scale in a large funding opportunity for consortia of states to develop college- and career-readiness assessments that were to be accessible to the greatest numbers of

students (Race to the Top Fund, 2009), formalizing a paradigm shift in thinking about accessibility (Larson et al., 2020). Rather than a paradigm in which accommodations were the only avenue to accessibility, accessibility became the guiding goal for all tested individuals, with tiers of support from universal design for all, to features designated for any student by an adult or decision-making team, to accommodations for students with disabilities and English learners.

It is because of the historical progression in thinking about accessibility that we first address fairness through accessibility for individuals with disabilities and then for English learners. Following these two groups, we explore how accessibility has been part of the vision for other groups, including racialized and ethnicized groups, and individuals from low socioeconomic backgrounds. There is support for such attention in the *Standards*:

For some test takers, factors related to individual characteristics such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or English language proficiency may restrict accessibility and thus interfere with the measurement of the construct(s) of interest. (AERA et al., 2014, p. 52)

However, we note that it is not racial stratification or other characteristics themselves that are the relevant factor, but the discrimination and racism that create disparities in outcomes; this is addressed more directly in the final section on accessibility through cultural and linguistic responsiveness.

Individuals With Disabilities

Early work focused on nonstandard administrations of higher education admissions testing (Laing & Farmer, 1984; Willingham et al., 1988) as the way to enable certain individuals with disabilities to participate in assessments. The researchers who examined nonstandard administrations of the ACT (Laing & Farmer, 1984) and the SAT and GRE (Willingham et al., 1988) focused on who might be eligible for accommodations and the effects of specific accommodations (e.g., statistical differences in predictive validity, reliability, factor structure, differential item functioning). During this early research, so-called nonstandard administrations (i.e., accommodations) included the use of braille, large type, and extended time.

Defining accessibility as the provision of accommodations continued for decades and early on expanded beyond college admissions exams to assessments administered to K–12 students with disabilities. Accommodations initially were defined as changes to the materials or procedures of testing that allowed students with disabilities to take tests but did not change the content or targeted knowledge and skills being measured (Pitoniak & Royer, 2001; Thurlow et al., 2003). These notions were later refined to mean changes that minimized construct-irrelevant variance and promoted the measurement of what students with disabilities know and can do, thus providing access while ensuring the validity of score interpretations and uses when accommodations are used (Bolt & Roach, 2009; Laitusis & Cook, 2007). In the *Standards*, test accommodations are

adjustments that do not alter the assessed construct that are applied to test presentation, environment, content, format (including response format), or administration conditions for particular test takers, and that are embedded within assessments or applied after the assessment is designed. Tests or assessments with such accommodations, and their scores, are said to be *accommodated*. Accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together. (AERA et al., 2014, p. 215)

Use of accommodations as a means to access for K–12 students was supported through policy changes, which resulted in part from analyses of state testing policies (Thurlow et al., 1993, 1995). The 1997 reauthorization of IDEA required that all students with disabilities participate in state- and district-wide assessments with accommodations as appropriate. This requirement heightened awareness of the role of accommodations during assessments and, in turn, their use during instruction and their implications for score interpretation and use (Heaney & Pullin, 1998). It also led, after considerable study, to the provision of accommodations in NAEP (Lawton, 1995; Olson & Goldstein, 1996).

Concern about access to the general education curriculum and assessments through the availability of accommodations in instruction and during testing has generated a rich body of research (e.g., Buzick & Stone, 2014; Cawthon & Leppo, 2013; Gregg & Nelson, 2012; Laitusis & Cook, 2007; Thurlow & Kopriva, 2015). Extensive work has improved the analyses of state assessment accommodations policies (e.g., Lazarus et al., 2009) and documentation of the percentage of students with individualized education program (IEP)-assigned accommodations (see Thurlow & Wu, 2019).

In 2004, the reauthorization of IDEA required states to report on the number of students with disabilities receiving accommodations during statewide assessments. These data revealed the large variation in the percentage of students with disabilities who are assigned accommodations across states (Thurlow & Wu, 2019), which, depending on the grade and subject area, could range from as little as 0.7% to as high as 89.9% of students with disabilities. They also provided information about differences in accommodations assigned for reading and mathematics tests and by grade level. Unfortunately, only state-level data for general education assessments are available; those data do not provide information on whether this type of variability exists for students who have the same disability category label.

Researchers, through hundreds of studies, have focused on a wide array of accommodations (e.g., read aloud, text to speech, sign language, word prediction) for these test takers and examined not only the effects of accommodations, but also the results of surveys about their use. A substantial body of research has been accumulated on the effectiveness of accommodations for individuals with disabilities. Phillips (1994), in presenting a case for defending the validity of accommodations, asked, “Would non-disabled examinees benefit if allowed the same accommodation?” (p. 104), where an affirmative answer would suggest that the accommodation may not be defensible. The differential-boost hypothesis (Fuchs et al., 2000) proposed that an accommodation

that increases accessibility for individuals with unique needs should result in higher scores for individuals with disabilities but not for individuals without disabilities. Later, this hypothesis was replaced with the interaction hypothesis (Sireci et al., 2005), which proposed that the increase in performance would be greater for individuals with disabilities than for other individuals, but that the scores of all test takers might increase some.

Several meta-analyses and reviews have been conducted to aggregate the effects of accommodations for individuals with disabilities (Buzick & Stone, 2014; Gregg & Nelson, 2012; Rogers et al., 2014, 2016, 2019, 2020). Although it was generally reported that accommodations do in fact make a difference in the performance of individuals with disabilities, those effects varied in complex ways across test subject matter, the nature of the accommodation, and the ways in which accommodations were administered.

The singular focus on accommodations for access and fairness changed in the 1990s when advocates realized that accommodations were not a sufficient mechanism for providing access for some students with disabilities. In 1997, IDEA required that states develop (by the year 2000) alternate assessments for those students who were unable to participate in the general assessment with or without accommodations. The evolution in the development of these alternate assessments covered years in which there was confusion about the performance standards of other alternate assessment options that had been allowed by policy and regulations, such as alternate assessments based on grade-level achievement standards (Quenemoen, 2009; Wiener, 2005) and alternate assessments based on modified achievement standards (AA-MAS; Lazarus et al., 2015). The AA-MAS generated concerns about lower (inequitable) levels of expectation (Thurlow, Lazarus, & Bechard, 2013) and psychometric quality (Rodriguez, 2009, 2011); eventually, the regulation allowing it was rescinded (U.S. Department of Education, 2015), just months before the enactment of the Every Student Succeeds Act (ESSA, reauthorization of ESEA), which disallowed any alternate assessment other than the alternate assessment aligned to alternate academic achievement standards (AA-AAAS) for ESSA accountability purposes.

ESSA clarified that the AA-AAAS is appropriate only for students with the most significant cognitive disabilities. Because *students with the most significant cognitive disabilities* is not an IDEA disability category, states are required to develop their own guidelines and definitions (Thurlow et al., 2019). Researchers have found, though, that students who participate in AA-AAAS tend to have intellectual disabilities, autism, and multiple disabilities (though not all students in those categories have the most significant cognitive disabilities), and yet still vary widely in their characteristics (Erickson & Geist, 2016; Erickson & Quick, 2017; Kearns et al., 2011; Towles-Reeves et al., 2009).

Further, ESSA clarified that students with the most significant cognitive disabilities are to be taught and assessed on the same content standards as their peers without disabilities, although possibly at reduced depth, breadth, and complexity. The test development community has been challenged by the need to develop AA-AAAS

appropriate for the range of students in this group, with the goal of producing results that support fair and valid test interpretations and uses. Despite these challenges, AA-AAAS are held to the same requirements as other tests for design and development that supports fair and valid test interpretations and uses (U.S. Department of Education, 2018). Much has been learned in the process of developing AA-AAAS that meet these requirements (e.g., Thurlow & Quenemoen, 2016), yet the expansion of what accessibility means for these students continues to be a challenge. The progression of thinking about accessibility for school-age students with disabilities leads to the argument that, for certain individuals, access could mean different standards for performance, in part because of the possibility of differences in disciplinary knowledge and skills covered in instructional practices and tailored approaches to learning to meet individualized learning goals, as well as extended learning periods through age 21 and beyond. This shift in thinking has not (yet) drifted into other assessments and other individuals being assessed.

As noted earlier, nonstandard administrations began in the 1980s with attention to accessibility for individuals with disabilities outside the K–12 school system. By 1990, the Americans With Disabilities Act required that any business or agency receiving federal funds had to make accommodations available. It stipulated

(A) making existing facilities used by employees readily accessible to and usable by individuals with disabilities; and (B) . . . acquisition or modification of equipment or devices, appropriate adjustment or modifications of examinations, training materials or policies, the provision of qualified readers or interpreters, and other similar accommodations for individual with disabilities. (42 U.S.C. 12/11, Section 101[9])

Clarification of the individuals who qualified for accommodations was provided in the 2008 reauthorization of the Americans With Disabilities Act; this clarification was viewed as broadening the population of individuals with disabilities who were eligible for accommodations in higher education and the workplace. Because of concerns about the provision of these accommodations in practice, the U.S. Government Accountability Office (USGAO, 2011) issued a report calling for improved federal enforcement of the right to testing accommodations. It brought renewed attention to the role of accommodations regardless of age, as well as to the decisions made about whether to provide accommodations to individuals:

Given the critical role that standardized tests play in making decisions on higher education admissions, licensure, and job placement, federal laws require that individuals with disabilities are able to access these tests in a manner that allows them to accurately demonstrate their skill level. While testing companies reported providing thousands of test takers with accommodations in the most recent testing year, test takers and disability advocates continue to raise questions about whether testing companies are complying with the law in making their determinations. (USGAO, 2011, p. 29)

Attention to college admissions testing continues to increase, in part because of state adoptions of college admissions tests since the NCLB act required high school state-wide assessment. This attention is to a broader array of individuals with disabilities and a broader set of accommodations. Although testing companies report providing thousands of test takers with accommodations, test takers and disability advocates continue to question whether testing companies are complying with the law in making their determinations (USGAO, 2011).

Including the K–12 and higher education systems, individuals with disabilities comprise about 13.6% of the population (2021 American Community Survey; U.S. Census Bureau, 2023). These individuals are, at various times in their lives, taking tests. For example, the number of individuals with disabilities taking the Praxis II tests for educators is increasing (U.S. Department of Education, 2010). Estimates of overall participation of individuals with disabilities in credentialing exams are not available (van den Heuvel et al., 2016).

Attention also has been given to considerations for individuals with disabilities taking other types of tests, particularly those used for career decisions (Nester, 1993), including law exams (Eichorn, 1997; Thurlow et al., 1997) and medical exams (Little, 1999), to name a few. These early works almost exclusively focused on a concern about giving test takers an unfair advantage by providing them with accommodations during testing (Phillips, 1994).

We acknowledge that early employment of accommodations was an important vehicle to provide access to tests for many individuals. However, in the spirit of fairness through access, these earlier attempts were not sufficient for many to accurately and meaningfully represent their disciplinary knowledge and skills in ways that were relevant and responsive to the purposes of the tests.

The addition of alternate assessments was one way to provide access to assessments for those students with more significant or challenging disabilities. Still, over time, this access avenue was limited by federal law to a very small percentage of the student population, thus limiting their potential as a general accessibility approach. In addition, universal design principles have received partial attention rather than full application and use in a way that would eliminate much of the need for accommodations. Part of the hesitation may be due to the need to document the access needs of students with disabilities, which are generally documented on IEPs as accommodations. Further, it is only recently that other avenues to access, such as computer-adaptive tests, have attempted to conform to a person's ability level by focusing item selection to optimally locate the person on the ability continuum (reducing the need to administer items that are too hard or too easy and thus less informative). A major limitation with computer-adaptive tests is that they select from items that are limited in scope, modality, and format, as constrained in the item and test specifications. As of yet, there are not computer-adaptive tests that either select or build tasks (through automated item generation or the use of artificial intelligence) that acknowledge how the test taker might best engage in an assessment of their content knowledge and skills.

Universal Design

Universal design is largely known in the assessment industry through descriptions in the *Standards* and publications of the National Center on Educational Outcomes (Thompson et al., 2004, 2005), referred to as UDA throughout this chapter. In the *Standards*, UDA is described as a means to maximize fairness by defining constructs precisely, avoiding item or test characteristics that may introduce construct-irrelevant variance specific to some groups, and being intentional about selection of font size, administration time, and linguistic load, among others. In addition, UDA calls on test developers to “avoid item contexts that would likely be unfamiliar to individuals because of their cultural background” (AERA et al., 2014, p. 58).

CAST was founded in 1984 and has championed UDL. Although they have undergone revision over time, CAST promotes UDL guidelines to secure access and participation of all learners in challenging and appropriate learning opportunities. The guidelines are based on the why, what, and how of learning, through the provision of multiple means of engagement (stimulating and motivating learning), representation (presenting content in multiple ways), and action and expression (differentiating ways for students to express their knowledge and skills). In 2020, CAST initiated a community-driven process to revise the UDL guidelines (<https://udlguidelines.cast.org/>), with attention to equity, inclusion, and fairness. These guidelines are much broader than those for UDA and explicitly enable culturally and linguistically responsive approaches to assessment through the acknowledgment of multiple ways (means) of knowing and doing (we address this directly in the final section).

Implications for Practice and Research

Test developers and users should employ the principles of UDA and UDL for test and assessment development and administration and, as a corollary, encourage the use of UDL to enhance accessibility of instruction and learning experiences. Test developers and researchers should support research programs addressing the effectiveness of the application of UDA and UDL, particularly regarding the extent to which more intensive applications of these principles might reduce requests for accommodations. In addition, state and federal education leaders should intensify efforts to develop high-quality AA-AAAS as a positive direction to achieve accessibility and acknowledge important differences in how some individuals learn, access disciplinary knowledge and skills, and convey what they know and can do.

English Learners

English learners emerged as a group that required special consideration for inclusion in K–12 education testing programs almost immediately after concerns were initially raised about the exclusion of students with disabilities (see August & Hakuta, 1994; LaCelle-Peterson & Rivera, 1994). In general, in the United States English learners are students whose native language is not English and who are eligible for English language development services. They have difficulties in reading, writing, speaking, or listening

in English that are severe enough to deny them the ability to achieve in classrooms where the language of instruction and assessment is English (National Center on Educational Outcomes, 2020). These characteristics result in difficulty accessing not only the content of instruction but also the assessments designed to measure their content knowledge and skills, as well as their English language proficiency.

In the K–12 education system, identified English learners are eligible to receive English language development services, with the expectation that eventually they will leave English learner status because they have gained proficiency in English reading, writing, speaking, and listening (Linguanti et al., 2013), thus being prepared to fully participate in English-instruction classrooms. Researchers have estimated a 3- to 7-year period to develop academic English proficiency to support schooling, with a slightly longer time for older immigrant English learners, with a shorter range of 3 to 5 years for oral proficiency and a longer time given socioeconomic background and age of immigration (Collier, 1989; Garcia, 2000; Hakuta et al., 2000).

English learners have been the subject of research on assessment accessibility. In part, this attention followed the attention given to students with disabilities, and much of the research has been carried out in the K–12 school system. Similar to individuals with disabilities, access to the curriculum and to assessments is a concern for English learners. Increasingly, researchers focused on the effects of various accommodations for English learners (e.g., Abedi & Hejri, 2004; Abedi et al., 2000, 2004; Kopriva et al., 2016; Kosak-Babuder et al., 2019). A substantial body of research has been accumulated on the effectiveness of accommodations, following a similar methodological path from a differential-boost hypothesis to an interaction hypothesis for these students. A number of meta-analyses and reviews also have been conducted to aggregate the effects of accommodations for English learners (Kieffer et al., 2009; Li, 2014; Liu et al., 2018; Vanchu-Orosco, 2012). Findings generally are less promising for English learners than for students with disabilities, but still vary in complex ways across test subject matter, the nature of the accommodation, and methods of accommodation administration. In a systematic review, Rios et al. (2020) reported that for the most part, accommodations used with English learners lacked evidence of effectiveness; most were accommodations employed for individuals with disabilities.

Researchers confirm that accessibility is a broad issue that applies beyond school years. With the National Education Longitudinal Study (National Center for Education Statistics [NCES], 1988), researchers revealed that 12 years after their eighth-grade year, only about 13% of English learners had earned a bachelor's degree, compared to 25% of non-English learners (Kanno & Cromley, 2013). Attention to the college-entrance assessments that these and other students take to gain access to postsecondary education is increasing (e.g., Moore et al., 2018), yet remains underdeveloped.

There is a growing body of work that supports the use of bilingual or multilingual assessments, in cases where English would be the language of instruction and assessment. Bilingual assessment can take multiple forms (Caesar, 2020). On a minimalist

side, with digital delivery of tests, the test developer can embed a translation dictionary in the test by enabling a hover-over dictionary tool (or some other process to get immediate translation of each word in a test). A more encompassing approach is to provide a bilingual experience, where items are presented in a native language and English simultaneously (e.g., side by side). Bilingual assessment embodies the characteristics of accessibility by opening up access to the test and test questions by eliminating the barriers due to language (e.g., see efforts with the National Assessment of Educational Progress; NCES, 2023).

Another approach is to create assessments in multiple languages, as done with the international assessment survey programs (e.g., Trends in International Mathematics and Science Study [TIMSS] of the International Association for the Evaluation of Educational Achievement), especially in contexts with non-English instruction, including in learning settings that are engaging in language revitalization and immersion. Multilingual assessment efforts require additional layers of psychometric analyses to ensure score comparability because the explicit intention is for international comparisons of disciplinary knowledge and skills. Testing programs develop and deploy common content frameworks that are used to develop test specifications through which items are developed and translated into dozens of languages.

When the intent is to provide optimal access, translation makes sense. But in the United States, where the primary language of instruction is English and although there is no official national language, most state testing programs function under an assumption that an important outcome of education is proficiency in English (implied by federal school accountability regulations). However, there are states that test in multiple languages at various grade levels in both reading and mathematics to accessibly measure disciplinary knowledge and skills without the interference of English language proficiency. At least one other state, Hawaii, has encompassed Indigenous language knowledge and practices into state achievement content standards and thus has included tasks relative to Indigenous language in the state achievement tests. There are a growing number of examples of the inclusion of Indigenous language knowledge and practices in national assessment programs across the globe (e.g., Guatemala, New Zealand, Canada, and Australia).

The National Council on Measurement in Education (NCME, 2020) statement on testing with English learners promotes an approach to test development and administration that acknowledges specific factors relevant to these learners. The statement emphasized consideration of the following: (a) language is culturally grounded because culture is often expressed through language; (b) when possible, the language of the test should be consistent with the language of instruction; (c) test administration modes should be familiar for English learners; (d) English receptive and productive language skills should be strong enough to provide full access to the test; and (e) the presence of special needs or disabilities further complicates test accessibility. With greater understanding of the nature of English language development, test developers will achieve greater fairness in testing for all students.

Faulkner-Bond (2020) argued for the consideration of the language of test takers relative to disciplinary language that may be embedded in the test. She recommended oversampling English learners in test development and validation processes to gain deeper understanding of the role of language uses in local settings, in instruction, and within disciplines, as well as discipline-specific language use in the articulation of content standards (how we describe what test takers are expected to know and be able to do). The fundamental ways test takers come to learn and understand disciplinary content and practices are influenced by local contexts, sociocultural contexts, instruction (as practiced locally), and other ways. The operational assumption that test scores are not influenced by language proficiency not only ignores these contexts, but also ignores that we often know less about the language proficiency of non-English learners than we do of English learners. After reviewing the structure and results of several assessments of English language proficiency, regarding comparability, Faulkner-Bond suggested that the focus should be on comparability of uses and interpretations rather than psychometrics, as well as on how proficiency is defined and measured.

Finally, there is growing recognition that native language skills are important to not only sustain, but also continually develop, particularly in early childhood (Duran et al., 2019; Wackerle-Hollman et al., 2019, 2022). In addition, most states award a seal of biliteracy

in recognition of students who have studied and attained proficiency in two or more languages by high school graduation . . . to help students recognize the value of their academic success and see the tangible benefits of being bilingual. (<https://sealofbiliteracy.org/>)

Sufficient evidence exists to support multilingual development, begging the question as to why so many tests continue to be administered in English only.

Implications for Practice and Research

Test developers and users should consider the role of language of instruction, language proficiencies in test and native languages, native and test language similarities, availability of validity evidence to support score interpretation and use, evidence regarding fairness, possible presence of specific disabilities, quality of English learner identification, and quality of language instruction and supports. They also should appreciate and recognize the sociocultural contexts in which students live and learn, from test and assessment design to score reporting, interpretation, and use. The research basis for the use of accommodations with English learners is inconsistent and limited, with stronger promise based on research with forms of bilingual assessment—an area that deserves additional attention, consistent with the goals of accessibility and UDL.

Racialized/Ethnicized Groups

Because of the persistence of disparities between racialized and ethnicized groups in education opportunities and outcomes, federal legislation required states to report

on the education achievement levels of major racialized or ethnicized groups. We acknowledge that these persistent disparities are largely driven by oppression and racism, limiting education access and opportunities. Because of this, federal legislation has been used as a tool to ensure education access and opportunities. As part of President L. B. Johnson's War on Poverty, the ESEA authorized the federal government to direct federal dollars to the most disadvantaged children through state governments, resulting in wide expansion of state departments of education. Through the 1994 reauthorization of ESEA as the Improving America's Schools Act, states were required to implement challenging content and performance standards and state assessments of them; to include all students in them; to report on performance broken down by racialized and ethnicized status, gender, English proficiency, migrant status, disability, and economic status; and to hold schools accountable. Because many states either did not recognize the requirements or chose to ignore them, the 2001 reauthorization of ESEA, the No Child Left Behind Act (NCLB), significantly increased the Title I accountability requirements with sanctions for failure to do so. States were required to establish grade-level content standards and assessments of achievement of those standards; to report performance (achievement proficiency rates) of students by racialized and ethnicized status, and socioeconomic status, as well as English learners and students in special education; and to hold schools accountable for adequate progress toward the standards (see Ho & Polikoff, this volume).

When investigating fairness and equity in education opportunities and outcomes, racialized categories have been employed for decades (Bohrnstedt et al., 2015; Coleman et al., 1966; Cottrell et al., 2015), with some evidence that standardized testing may play a role in reinforcing segregation (Knoester & Au, 2015). In this work, the concept of race implies a history of housing segregation (resulting in education and occupation disparities) that is connected to occupation segregation (resulting in income and health disparities) that is connected to education segregation (resulting in occupation disparities)—cycles of segregation that limit access to education and career and life opportunities (Cottrell et al., 2015). We note that these cycles are the product of oppression and racism that further the reproduction of disparate opportunities and outcomes. Race and ethnicity are relevant to the extent that they reflect different levels of performance, as an outcome of racism, and such differences call into question qualities of the test itself or result in exacerbating existing disparities and inequality through unintended negative consequences. For example, in 1969, the Association of Black Psychologists called for a moratorium on testing Black children because of disproportionate placement in special education and developmental settings (a result of discrimination and racism), further excluding them from general education opportunities (Garrett Holliday, 2009).

Perhaps the most significant regulations regarding the role of racialized status in education outcomes came in the NCLB reauthorization of ESEA in 2001. President George W. Bush called on the country to end the soft bigotry of low expectations,

acknowledging that achievement gaps have persistently fallen along racial and socioeconomic lines. However, Bush mistakenly argued that education achievement gaps produced discrimination, arguing for a causal direction that was inconsistent with the vast body of evidence indicating the opposite (Rubel & McCloskey, 2019). Through the rules and regulations from NCLB, states were required to report student group performance on academic standards-based tests based on racialized and ethnified status, socioeconomic status, English learner status, and special education status. This was the first federal requirement for states to report on and respond to student group performance that was rigorously enforced, with a requirement that to be counted as achieving adequate yearly progress, 95% of students had to participate in the state's assessments. Prior to this, some students often were exempted from participation because of various exceptionalities, particularly for students receiving special education services (McDonnell & McLaughlin, 1997) or English language services (August & Hakuta, 1997).

Substantial research has addressed the associations between concentrated poverty and segregation in schools with disparities in education supports, lower expectations, increased disciplinary actions, and education achievement and school completion. Even though the nation's diversity continues to increase, schools in the United States are becoming increasingly more segregated, with some regions returning to pre-civil rights levels of segregation (Frankenberg et al., 2019; Reardon et al., 2019); again, we note that racial segregation is a direct result of oppression, discrimination, and racism. Using data from NAEP, it is clear that school achievement test performance is strongly associated with school composition or segregation (Bohrnstedt et al., 2015), where in the United States,

- on average, White students attended schools that were 9% Black, whereas Black students attended schools that were 48% Black;
- schools with greater than 60% Black students tended to be located in urban areas and in the South and the Midwest;
- achievement was lower for Black and White students in schools that had the highest Black density but achievement gaps were not different regarding levels of Black density of schools; and
- the magnitude of achievement gaps, given Black student density of schools, was smaller when accounting for socioeconomic status (SES) and other characteristics, but was still nonignorable.

Many of the challenges that exist regarding racial stratification and educational testing are a function of inappropriate test interpretation and use, which are too often intertwined with discrimination and racism. We acknowledge that, in most settings, test use can provide a deeper understanding of individual and group knowledge, skills, and abilities and inform nearly all education decision-making. However, to realize the potential advances brought about by the use of educational tests, fairness must be secured. Part of the challenges that exist regarding racial stratification and educational testing are the

continued limitations of accessibility, particularly those associated with limited cultural and linguistic relevance of tests to culturally and linguistically diverse communities (this topic is discussed more fully in the section “Accessibility Through Cultural and Linguistic Responsiveness”).

Implications for Practice and Research

Test developers and users should secure the highest levels of scrutiny of test score quality and fairness. However, our current approach to test development and analysis is limited in the accessibility sense because of the exclusion of considerations for cultural and linguistic backgrounds of intended test takers. Education disparities have a deep and long history rooted in racism, which results in segregation and cycles of poverty. Federal requirements to report achievement levels by race and ethnicity uncover these disparities to some extent, and tests such as NAEP have been important indicators of the nation’s progress in addressing achievement disparities and the opportunity gaps producing them. To eliminate large-scale testing because it reflects disparities in opportunity would eliminate an important public policy indicator. However, to use such tests without acknowledging the contexts of racism and racial segregation could result in decisions that maintain segregation and continue to limit access to opportunities.

Socioeconomic Disadvantage

As with racial stratification, disparities in education opportunities and outcomes for students with different levels of SES have been a persistent challenge. SES also has a long-lasting trend connected to segregation and multigenerational poverty (Sharkey & Elwert, 2012). The authorization of ESEA was primarily focused on eliminating poverty through increased education opportunities, and reauthorizations through the Improving America’s Schools Act, NCLB, and ESSA maintained the focus on students and schools that experienced persistent disadvantage.

James Popham has long argued for the instructional relevance of testing and assessment. In his work on instructional sensitivity of tests, he argued that instructional insensitivity could result from items that produced correlations between item response and SES, where wealthier students would be more likely to respond correctly (Popham & Ryan, 2012). Popham (2007) argued that the SES composition of schools explained more variance in student performance on accountability tests than the effectiveness with which students have been taught or teachers’ instructional efforts. This position is consistent with decades of research where researchers reported that most of the variance in student achievement scores exists within schools rather than between schools, and most of the between-school variance is explained by school demographics resulting from segregation and not indicators of school quality (Coleman et al., 1966; Rodriguez & Nickodem, 2018).

We also acknowledge that SES and racial stratification are highly confounded because there exist significant disparities in wealth as a function of racial stratification, a trend

that continues to grow, as has the wealth gap between the richest and poorest families in the United States (Schaeffer, 2020). SES is clearly a construct-irrelevant factor, yet to the extent that it explains variation in access to opportunities and resources, it explains variation in achievement. Once again, elements of segregation and oppression are intertwined with cycles of poverty and limited access to opportunities that may provide greater education and social mobility.

Test Takers Facing Multiple Access Needs

Aspects of test development and administration that limit accessibility to some individuals with certain characteristics associated with education inequity and disparities tend to affect performance for multiple reasons, in part because of the multilayered and intersectional nature of disparities resulting from discrimination and oppression. For example, in 2015–2016, the percentage of students served through IDEA (special education services) was highest among American Indian/Alaskan Native students (17%) and Black students (16%) relative to White students (14%; NCES, 2019). Similarly, in 2016, 24% of Black and Latino children under the age of 18 were in families at or below the poverty level (based on the supplemental poverty measure), relative to 8% of White children (NCES, 2019). In addition, in fall 2017, 45% of the nation's Black and Latino students attended high-poverty schools (more than 75% eligible for free or reduced-priced lunch), whereas 8% of White students did so; 31% of White students attended mid- to high-poverty schools, whereas 75% of Black and Latino students did so (NCES, 2021). Note the differences between poverty rates and attendance in high-poverty schools as another indicator of segregation. There are persistent disparities in SES, employment, income, stable housing, access to healthcare, and other economic indicators across racialized and ethnicized groups, all of which have implications for education opportunities and outcomes.

Although much of the discussion about fairness for specific groups of test takers focuses on one group at a time, in reality, most of the individuals in these groups also belong to other groups for which there are concerns about fairness. For example, individuals with disabilities are frequently among the group of individuals in poverty; the U.S. Census Bureau reported that the poverty rate of individuals with disabilities aged 18–64 in 2018 was 26.9%, compared to 9.5% of individuals without disabilities (Semega et al., 2020); similarly, the U.S. Department of Labor (2021) noted that “across all age groups, persons with disabilities were much less likely to be employed than those with no disabilities” (p. 1). Among school-age children, researchers also have noted the connection between disability status and some disability categories with poverty (e.g., Schifter et al., 2019). Similarly, being an English learner is often associated with poverty and may be associated with immigration status.

Interest continues to grow about how best to assess the school-age population of English learners who have disabilities because of the unique characteristics they present in achieving a fair assessment process, characteristics that most testing programs are not designed to adequately address. The population of school-age English learners with

disabilities is increasing at a higher rate than the population of students with disabilities overall (Wu et al., 2021). Historically, the access needs of these students have been addressed in one of two ways. First, these students often have been treated as having special education needs in ways that eclipsed their need for English language development services and accessibility supports (National Council on Disability, 2018). This meant that when they were assessed, consideration was given to the disability access needs but not to their limited English needs. Second, as the access needs of English learners have been recognized, educators have been asked to think of a student's needs as though the student were two individuals—addressing the disability needs first, then addressing the English language needs; if there were conflicts in accessibility supports, then, again, their disability needs would be treated as more important than their English language development needs.

For too long, attention to the access needs of English learners with disabilities only focused on those English learners who did not have significant cognitive disabilities (National Center on Educational Outcomes, 2014). Questions about how to provide assessments that meet their access needs have been raised, with little action until federal law required that these students participate in both alternate achievement assessments and English language proficiency assessments (Christensen et al., 2018; Thurlow et al., 2017; Winter et al., 2018).

Educational measurement practices regarding accessibility for students with multiple categories of access needs continue to improve. Most states provide assessment accessibility manuals that address both students with disabilities and English learners, consistent with the recommended approaches of the Council of Chief State School Officers (Lazarus et al., 2021). However, these efforts are less likely to be occurring beyond testing in K–12 education settings, such as higher education admissions assessments (Lazarus & Thurlow, 2016) and certification and licensure assessments (Lazarus et al., 2017).

Implications for Practice

Test developers and users should analyze the role of disabilities, English learner status, racial stratification, living in poverty, and other public policy relevant characteristics, simultaneously or in combination to obtain a clear picture of item and test performance in complex social contexts. For example, they could employ or create methods to disentangle certain learning disabilities and English learner status because the natural process of being a multilingual learner may present behaviors that appear to indicate language delays. Additional recommendations are provided in the following sections regarding the role of differential item functioning and measurement invariance analyses. In the prevailing empirical research paradigm, we test interaction terms before testing and interpreting main effects. If the association between X and Y depends on Z , then the main effect of X is misinterpreted without considering the level of Z . In some cases, the multiple effects of X and Z on Y are additive or even multiplicative, and possibly nonlinear.

REALIZING FAIRNESS THROUGH TEST DEVELOPMENT

Test Development

Fairness finds its introduction in the earliest test development stages, from conceptualization to administration. Principled assessment design (PAD; see Huff et al., this volume) provides a flexible and explicit comprehensive approach to test development that acknowledges the important roles of UDA and fairness principles. As described by Huff et al., PAD requires the definition of the construct or domain to be based on disciplinary learning science evidence of how individuals learn and develop disciplinary knowledge and skills. Second, PAD requires “the explicit articulation of all assumptions, design decisions, and rationales for those decisions” (p. 446) through the employment of a coherent set of design tools. Third, PAD also makes explicit the reality that assessment is a process of reasoning from imperfect evidence, but when the evidence is grounded in principled design, tests can meet their most challenging purposes. Huff and colleagues acknowledge the relevance and connections between accessibility and culturally and linguistically responsive approaches to test development, noting that “without the precision and transparency demanded by the PAD process, incorporating accessibility and culturally and linguistically responsive features could jeopardize the validity of the inferences about what students know and can do” (p. 494). The authors also argue that evaluation of cultural and linguistic responsiveness is not done one item at a time, but across an item pool or an entire test form. In addition, accessibility and cultural and linguistic responsiveness are associated with motivation and engagement, which further enhance accessibility for all students.

Embedded in PAD are principles of UDA (Ketterlin-Geller et al., 2015; Thompson et al., 2002, 2004), which require test developers to fully consider and embrace all possible test takers during each stage of test development, beginning with the clarity of the construct definition or domain. Particularly when considering the purposes driving the development of a test and the construct definition stages, the extent to which purpose and construct definition apply consistently to all test takers will support fairness. Principles of accessible measurement (Beddow, 2012) also require greater attention to construct definition, to prevent many instances of differential item functioning and promote measurement invariance. In addition, while addressing the assessment of achievement with a focus on classroom assessment, Shepard et al. (2018a) reminded us of the importance of context:

All learning is fundamentally social, involving the individual's use of shared language, tools, norms and practices in interaction with his or her social context. . . . Sociocultural theory offers a powerful, integrative account of how motivational aspects of learning—such as self-regulation, self-efficacy, sense of belonging, and identity—are completely entwined with cognitive development. (p. 23)

This position has implications for domain specification and test development. Construct or domain definitions are often underdeveloped (naturally limited because of the lack of inclusion of diverse developers or narrow specification of disciplinary knowledge and practices), from classroom and state achievement tests to professional certification

assessments. In their argument centering on the role of sociocultural learning theory as a way to bring coherence to curriculum, instruction, and assessment, Shepard et al. (2018a, 2018b) argued that assessments designed to facilitate inferences about learning must be constructed with research-based models or theories of learning (the first component of PAD).

Regarding test development, fairness can be improved through explicit attention to fairness guidelines (Zieky, 2016) and quality assurance procedures (Allalouf, 2007, 2011, 2017) in the professional practices that produce test blueprints and specifications, item specifications, item-writing and -editing guidelines, item-writer training, item review processes, test assembly and review, scoring rules and computational procedures, and item and test analyses procedures of field-test and operational administration results (Oliveri & von Davier, 2016). Item-writing guidelines exist that specifically promote greater accessibility and fairness (Haladyna & Rodriguez, 2013; Rodriguez, 2011; Rodriguez et al., 2014; Zieky, 2016; see also ETS, 2022). Given the many approaches and stages of test development, the process should be a collaborative effort, rather than a linear staged process, through which fairness practices are shared among individuals with diverse experiences and expertise (Rodriguez, 2016).

As an example of accessibility-focused test development (or redesign), Wolfe and Gitomer (2001) demonstrated how domains can be made more accessible through improved assessment design and scoring. They focused their work on the National Board for Professional Teaching Standards' certification assessment of accomplished teachers, which includes a portfolio component where candidates provide evidence of accomplished practice, as defined in the standards. To do this, they set out to simply reduce the guessing candidates often engage in, trying to figure out (guess) what will be scored and what makes a difference. By explicitly describing how responses are scored and providing guidance to optimize decisions regarding the construction and content of the portfolio, they brought depth and clarity to the domains being assessed. In addition, they introduced additional structure to the assessment prompts and improved scoring by allowing for more benchmarks and training samples, rater training on bias, and rubric improvements.

Implications for Practice

Test developers and users should employ the principles of UDA and UDL through the PAD approach, perhaps in ways that go beyond surface features of tests and address the core aspects of domain specifications and inclusion of culturally and linguistically relevant disciplinary knowledge and practice (see final section). PAD holds great promise as a test development process that centers accessibility; accessibility features in PAD should be amplified.

Innovations in Item and Task Development

With widespread use of computerized testing and remote or cloud-based delivery, test developers have explored many alternatives for item formats. We use the term technology enhanced (TE) to describe such items, but other terms essentially describe

the same notions, including innovative and interactive items. In part, the purposes of TE item development are to create opportunities for test takers to interact with items and tasks to deepen the domain representation, tap difficult-to-measure higher order cognitive processes, represent pedagogically relevant features, include multimedia components expanding the sources of stimulus materials and reduce text dependence, employ more authentic contextualized elements in the test, allow for a wide variety of response modes, and increase test-taker motivation and engagement. These not only are goals of item development, but also become components or assumptions underlying the interpretation of item responses and test performance and introduce questions of accessibility and fairness.

There are too many forms of TE items to create an exhaustive list (see Bennett et al., this volume). However, some TE item features have become more commonplace in large-scale testing, including the use of drag and drop (structured or freeform), hot spots (clickable regions on images and displays or among listed options), figural drawing, ordering and sorting, and the use of a wide range of interactive tools such as balance beam, measuring devices, thermometer, telescope, compass, and even science lab equipment (virtual simulations). Russell and Moncaleano (2019) reported on the prevalence of various item types. They found that approximately 40% of 236 TE items that were released by large-scale K–12 education test programs aligned with a high level of domain fidelity, whereas nearly 40% provided low domain fidelity. This indicated potential challenges of TE items to meet the construct and domain definition clarity required by UDL and PAD.

One tool to support these efforts is the Technology-Enhanced Item Utility Framework (Russell, 2016; Russell & Moncaleano, 2019), designed to examine the fidelity with which TE items represent the intended construct. The framework enables TE item evaluation through structured human judgment regarding usability and accessibility, as well as construct fidelity.

Validity research has lagged behind the advances in item and test delivery innovations, in part because item developers often create tasks and sometimes deploy them without sufficient validation beyond psychometric item analysis. Early investigations of TE items addressed questions of psychometric properties and efficiency (Jodoin, 2003; Zenisky & Sireci, 2002). Some efforts have been successful in identifying TE items that may or may not contribute to domain representation and support test score interpretation. This work includes, as examples, think-aloud investigations of mathematics and English language arts items (Dolan et al., 2011), examination of associations with other measures and predictive evidence for a carpal tunnel release surgery simulation (Shanedling et al., 2010), and the detection of construct-irrelevant variance in a test involving the creation of mathematical expressions (Gallagher et al., 2002).

There are critical fairness and accessibility issues that must be addressed in the design and use of TE items (Stone et al., 2015; Strain-Seymour et al., 2009), such as accessibility for individuals with vision and fine-motor disabilities. Consider, for example, drag-and-drop items—this item type requires visual, fine-motor, and hand–eye coordination skills. Most TE item formats have similar requirements, requiring the test taker

to interact with elements presented on screen, thus presenting significant accessibility challenges. A team of researchers in the Accessibility for Technology-Enhanced Assessments Project (Shaftel et al., 2015) spent 3 years investigating questions of accessibility. Employing the principles of evidence-centered design (Mislevy & Haertel, 2006; Huff et al., this volume) and UDA (Thompson et al., 2002, 2004), including UDL for technology-enhanced assessments (Dolan et al., 2013), they engaged a panel of vision and motor skill experts to consider several categories of processing: perceptual (vision, hearing, sensory modalities), linguistic (oral, written, sign, braille), cognitive, motoric, executive (engagement, attention, motivation), and affective (test-taker psychological states and moods). The researchers studied TE items in original and accessible formats, engaged students in cognitive labs, and field tested items in 11 states.

The team found that “item accessibility can be improved for all students on all item types by providing access to assistive technologies” (Shaftel et al., 2015, p. 128). They recommended that (a) instructions should explicitly declare what test takers are expected to do; (b) graphic and font features should be visually simple and large and avoid dependence on colors; and (c) tasks requiring fine-motor skills, such as scrolling, should be minimized. “These studies permit preliminary conclusions about the effectiveness and fairness of alternative item presentations” (Shaftel et al., 2015, p. 165) for students with disabilities.

Researchers addressing the accessibility challenges of TE items illustrate that cognitive complexity is not fixed and depth of knowledge interacts with specific abilities of test takers. Braille and print versions of TE items may not retain the intended cognitive demands. Similarly, as noted earlier, familiarity with item formats and response modes is an essential element for fairness and accessibility, and this may differentially impact English learners and others with limited or interrupted formal education.

Implications for Practice and Research

Test developers should secure collaboration among item writers, accessibility experts, subject matter experts, diversity specialists, psychometricians, and others, since this is essential in the design of TE items. They should engage in rigorous evaluation of item and test quality, with explicit examination of accessibility and fairness, with a focus on the extent to which the expansion of TE items and other innovations obtain responses relative to targeted knowledge and skills. Decisions about item format choice naturally have fairness implications, including standard (non-TE) item formats—test developers choosing among item formats must consider principles of fairness and access (Albano & Rodriguez, 2018; Rodriguez, 2002).

A Sociocultural Framing of Differential Item Functioning

As a standard step toward fairness, in the field-test or pilot stages of item and test development and periodically with operational data, differential item functioning (DIF) is conducted to identify items for bias and sensitivity review. The fundamentals of DIF and the uses of DIF in fairness are described by Zwick (this volume). We also

acknowledge that many existing DIF practices are based on test-taker groups with heterogeneous groupings, typically based on racialized and ethnicized status, English learner status, and special education status. Within each of these groups, there may be substantial heterogeneity that masks more important DIF results based on national original, language group or linguistic dialect group, and specific types of disabilities (Ercikan & Oliveri, 2013), as well as the needs described in the section “Test Takers Facing Multiple Access Needs.” Such efforts are typically limited because of the smaller, more specific groups of test takers resulting in samples sizes too small to support DIF methods. However, empirical approaches to identifying sources of DIF help protect us against ad hoc interpretations (Alavi & Karami, 2010; Karami & Salmani Nodoushan, 2011).

For a deeper treatment of DIF in the context of intersectionality, see Russell et al. (2022). Intersectionality provides a theoretical framework (Crenshaw, 1989), grounded within Black feminist theory, that addresses the intersections of racialized groups, gender, and other social identities, focused on the social positions within a hierarchy of social power—jointly shaping human experience (Bauer et al., 2021). In addition, Russell (2024) provided a review of DIF procedures in the context of multiple conceptualizations of justice.

Here, we explore the sociocultural context of DIF. To do so, we review three DIF studies that addressed the content and contexts of items, relative to test-taker characteristics. Social models of learning acknowledge that individual learning and cognition occurs through social interaction. Although focused on classroom assessment, Penuel and Shepard (2016) provided a summary of social and sociocultural models they deemed appropriate in the context of assessment. Sociocultural perspectives of learning extend social models into cultural spaces and acknowledge what cultural and linguistic traditions, knowledge, and practices test takers bring to the testing space (Penuel & Shepard, 2016).

An Example in Primary Education

Banks (2006) employed two culturally grounded taxonomies regarding Latino and Black cultures in the United States to identify common elements through which cultural bias could be examined in tests, while acknowledging the within-culture heterogeneity. She identified items from several large-scale tests that exemplified cultural elements. Using the Terra Nova test performance data from Latino, Black, and White fifth-grade students, Banks bundled items based on the presence of culture-specific elements in correct or incorrect options and evaluated for differential bundle functioning (correct-option culture-based item) and differential distractor functioning (incorrect-option culture-based item). Banks reported that a small number of items illustrating each cultural element limited statistical detection power. However, Banks did conclude that DIF was more likely to occur for distractors than for correct options. She also demonstrated how “cultural aspects illustrated in correct and incorrect options” (p. 131) may influence item functioning. She encouraged researchers and test developers to create more comprehensive means for classifying the cultural features of items.

An Example in Secondary Education

Stricker and Emmerich (1999) investigated the Advanced Placement (AP) Psychology examination and potential gender DIF. In their review of prior DIF research, they summarized potential sources of DIF, including familiarity with the item content, including exposure, experience, and cultural loading, as well as test-taker interest in the content and potential emotional reactions to item content. To further study such potential sources, the authors asked 717 students in 19 high school classes to rate items from the AP Psychology exam on familiarity, interest, and unpleasant affect with four-point rating scales (*not at all* to *very*). The standardized mean difference between females and males (d) was computed for each rating variable of each item and compared to Mantel–Haenszel values across items. Across the 13 content areas, gender differences in ratings for familiarity and unpleasantness were not significantly different. There were gender differences in interest ratings, where motivation, developmental psychology, and abnormal psychology were of more interest to females and biological psychology and social psychology were of more interest to males. However, these mean d effects for familiarity, interest, and unpleasantness were associated with Mantel–Haenszel means across content categories. The more that familiar and interesting items were in a content category to females relative to males, the easier the items were for females (DIF advantaging females); the more that unpleasant items were in a content category, the more difficult the items were for females (DIF disadvantaging females). The authors argued that the gender-based differences found in the AP Psychology items likely developed through differential socialization of females and males at home, at school, and in the community.

An Example in Postsecondary Education

O'Neill et al. (1993) examined the GMAT regarding the potential Black/White and female/male DIF relative to the content and characteristics of the items. Although the proportion of items with moderate to large DIF was low, they found interesting results. Regarding verbal items, they found that reading comprehension items with stimulus materials referring to Black-membered and Latino-membered people (*membered* denoting the socially derived identification of people) and those with social science content (acknowledging overlap in these characteristics) were easier for Black test takers than for the matched group of White test takers, with the opposite being true for general material (material with no reference to specific ethnicized contexts) and items with humanities and science content. Regarding quantitative items, Black test takers performed better on items focused on calculations or formulas, whereas White test takers performed better on word items requiring test takers to set up the problem and translate the words to numerical expressions to solve. In addition, Black test takers performed worse on items with business-related contexts than the matched White test takers. Black test takers also performed worse on items with 50 words or more in the stem/stimulus, whereas Black test takers performed better on items with fewer than 50 words in the stem than the matched White test takers. The authors

suggested the possible effects of differences in reading ability and the relevance of item characteristics and the enacted curriculum experienced by test takers. They noted that internal and external advisory committees should review such results to identify implications for test content.

O'Neill and McPeck (1993) acknowledged the relevance of social issues regarding DIF interpretation. They commented on differential opportunities afforded to students of color by less-resourced schools and differential treatment by teachers, as well as lower expectations for students of color, "even when all students are enrolled in the same class. Until we recognize that these education inequities exist, we will not be able to understand DIF results in the proper context" (O'Neill & McPeck, 1993, p. 276).

Implications for Practice and Research

Test developers and users should provide greater attention to the sociocultural context in which tests purport to measure academic achievement to enhance fairness with respect to accessibility, where such achievement is also embedded in sociocultural contexts (see Ercikan & Solano-Flores, this volume). An important function of DIF analyses is gained through the aggregation of lessons learned, lessons that should be turned into item and test development guidance. Test developers should design DIF analyses in ways that address the heterogeneity within groups and the relevant features of items and tests and supplement the typical item-by-item DIF flagging procedures undertaken by most testing programs. They should employ DIF analyses to achieve justice-oriented goals (Russell, 2024).

ACCESSIBILITY THROUGH CULTURAL AND LINGUISTIC RESPONSIVENESS

Given the deepening understanding of the role of sociocultural contexts in teaching and learning, and in testing and assessment (Shepard et al., 2018a), we support the use of culturally and linguistically responsive (CLR) assessment to increase accessibility.

The Role of Cultural and Linguistic Backgrounds

Accessibility can best be understood by contrasting the knowledge, skills, and abilities that reflect the construct(s) the test is intended to measure with the knowledge, skills, and abilities that are not the target of the test but are required to respond to the test tasks or test items. For some test takers, factors related to individual characteristics such as age, race, ethnicity, socioeconomic status, cultural background, disability, and/or English language proficiency may restrict accessibility and thus interfere with the measurement of the construct(s) of interest. (AERA et al., 2014, p. 52)

The first point is what Ketterlin-Geller (2008) would call differentiating target skills from access skills—although such skills may be interdependent, particularly when the

construct of interest is imbedded in real-world scenarios. The second point suggests that our humanity (individual and group characteristics and lived experiences) may restrict accessibility to the extent the construct of interest may reflect the humanity of some and not others.

The authors of the *Standards* offered general examples of how tests may limit access to the construct for some by including idiomatic phrases and regional vocabulary unrelated to the target construct or stimulus contexts unfamiliar to test takers given their cultural background (AERA et al., 2014, pp. 52–53). These test characteristics were not further explained and specific examples were not offered. Most test developers now do consider person characteristics such as, for example, blindness, dyslexia, and limited English proficiency, and do consider the potential barriers facing these test takers. These characteristics are in part the basis for UDA principles to improve accessibility, but they are limited and do not explicitly require test developers to consider the socio-cultural contexts of test takers.

A threat to fairness is in test content, vis-à-vis test-taker culture and linguistic histories. As an example, the authors of the *Standards* argued that critical reading

should not include words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups, or geographical locations, so as to maximize the measurement of the construct (the ability to read critically) and to minimize confounding of this measurement with prior knowledge and experience that are likely to advantage, or disadvantage, test takers from particular subgroups. (AERA et al., 2014, p. 54)

The authors continued:

Differential engagement and motivational value may also be factors in exacerbating construct-irrelevant components of content. Material that is likely to be differentially interesting should be balanced to appeal broadly to the full range of the targeted testing population (except where the interest level is part of the construct being measured). In testing, such balance extends to representation of individuals from a variety of subgroups within the test content itself. For example, applied problems can feature children and families from different racial/ethnic, socioeconomic and language groups. (p. 55)

We contend, as do the scholars and measurement specialists exploring CLR, that such a balance is contradictory with the earlier advice of what to avoid and destroys the value of measures of critical reading, when the readings themselves are based on narrowly defined content domains void of the cultural and linguistic realities of students. How is it possible to “feature children and families from different racial/ethnic, socioeconomic and language groups” (AERA et al., 2014, p. 55) while avoiding “words and expressions especially associated with particular occupations, disciplines, cultural backgrounds, socioeconomic status, racial/ethnic groups” (p. 54), and other contexts that make representation of different groups authentic? Giving

extensive guidance about the avoidance of sensitive materials to item writers results in little to no representation of diversity, and when representation is present, it is only in surface features of items, resulting in items and context material with diverse names and places only. The neutral content approach to item and test development is insufficient.

In fact, Randall (2021, 2023) argued that neutralizing items by eliminating any reference or expressions associated with cultural backgrounds or other contexts specific to some test takers places the item within the dominant cultural context, essentially White, middle-class contexts; neutrality is unattainable. Randall (2021) asked test developers to develop a commitment to justice, consider their own positionality, and consider the sociocultural and other identities of the test audience, including the voices of the intended test audience in the test development process.

Finally, the authors of the *Standards* presented opportunity to learn as a relevant context factor, which is “the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test” (AERA et al., 2014, p. 56). Furthermore, disparities in school resources in some settings, including those with large populations of linguistically, racially, and ethnically diverse communities, and possibly rural and isolated communities, affect the quality and content of teaching and learning. Opportunity to learn is a restricted concept because it presumes the target content domain is appropriate and (exhaustively) comprehensive. Does the (specified) content domain include cultural and linguistic disciplinary content knowledge and practices? If we fully adopted the principles of UDL, we would acknowledge many ways of knowing and doing through providing multiple means of engagement, representation, and action and expression in instruction and assessment.

To extend the discussion in the *Standards*, Standard 3.2 requires test developers to minimize the possible effects of “construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics” (AERA et al., 2014, p. 64). Much of the attention to such effects focuses on evidence such as DIF. However, when the entire test presents a uniform barrier, DIF will not uncover the effect on one or more items because the item performance is conditioned on the remaining total test score (which could be uniformly depressed as a result of a narrowly specified content domain). Similarly, a number of standards address the language of the test, relative to the native language or language proficiency of the test takers. Although a test may be administered in a relevant and appropriate language given the test purpose, as required in Standard 3.13, if the content is not relevant and appropriate, significant barriers to access remain. Another common method to address these barriers is external sensitivity reviews of test specifications and items—which typically focus on what items contain that may be problematic, not what items are missing in terms of CLR. Even though nearly every large-scale testing program has sensitivity review processes in place (many employ multiple stages of review), questionable items do find their way to operational tests (Dee & Domingue, 2021).

Content relevance is also addressed in the *Standards* regarding workplace testing and credentialing. An example is given where

“salt is to pepper” may be the correct answer to the analogy item “white is to black” in a culture where people ordinarily use black pepper, but the item would have a different meaning in a culture where white pepper is the norm. (AERA et al., 2014, p. 181)

In the context of developing educational tests, the authors of the *Standards* asserted that

focus is placed on measuring the knowledge, skills, and abilities of all examinees in the intended population without introducing any advantages or disadvantages because of individual characteristics (e.g., age, culture, disability, gender, language, race/ethnicity) that are irrelevant to the construct the test is intended to measure. (p. 187)

Standard 12.3 requires test developers and users of educational assessments to promote access to the construct for all intended test takers throughout all steps of test design. However, at this point, there is no discussion about the nature of the content reflected by test specifications, other than an acknowledgment that to allow access to test content, context, and response formats, test accommodations and modifications may be needed—again, a response to what is there, not an acknowledgment of what might be missing.

Although the authors of the *Standards* do not employ the language “culturally and linguistically responsive assessment,” the notions and elements of such an approach are present throughout. Authors of the “Fairness in Testing” chapter in the *Standards* interpreted fairness as “responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses” (AERA et al., 2014, p. 50), noting that such a definition exceeds what may be required legally. Similarly, they noted that strategies (accommodations) can be employed with tests and testing procedures “to be responsive to the needs of test takers with disabilities and those with diverse linguistic and cultural backgrounds” (p. 60) as well as other groups. Item and test developers are left with a challenge: to be responsive to the diversity of test takers regarding cultural and linguistic histories, but to avoid culturally and linguistically specific words and expressions and contexts that may not be familiar to some.

CLR Construct and Domain Definitions

Decisions about what to measure and how to measure pervade educational measurement and have done so since its beginning. Such decisions are often made not simultaneously, but sequentially, and not necessarily in a particular order. In education settings, many data collection strategies often default to a test as the tool of choice (how to measure), without substantial regard for the construct or domain of interest (what to measure). Certainly, some constructs and domains are better assessed through other means than a test. Such definitions are often made by policy boards (e.g., state school

board or professional associations), without the inclusion of broad representation from the test-taker population or their advocates or measurement specialists.

As described in the test design and development chapter in the *Standards*, the association between the test content and established content standards is critical—that “content specifications must clearly describe the content and/or cognitive categories to be covered so that evidence of the alignment of the test questions to these categories can be gathered” (AERA et al., 2014, p. 75). For most purposes, test specifications include descriptions of how the construct or content domain is represented on the test, which are often refined through the test development process (see Huff et al., this volume). Test specifications are defined broadly, including documentation of purpose and comprehensive details about content decisions. The test content should be fully described in the content domain, content specifications, or content frameworks. These often come from content standards defined by professional associations, education authorities (informed by subject matter experts), and job task or requirements analyses (see Margolis et al., this volume). Content decisions detailed in test specifications should include the sources and characteristics of reading passages, stimulus materials (e.g., illustrations, figures, graphs), practice-based cases, scenarios, vignettes, and contexts related to diversity and local, regional, or national relevance (Rodriguez, 2016).

In education settings, questions have been raised about the appropriateness or fairness of domain specifications and the identification of disciplinary ways of knowing and doing. Some of the clearest messages regarding fairness of construct or domain definitions come from the work of Indigenous educators. Tests should reflect relevant content and modes of understanding and learning to adequately assess what we expect individuals to know and be able to do. Although education researchers have long argued that culture and society play a role in cognitive development (e.g., Vygotsky, 1978), this acknowledgment has been late to influence assessment (for an example in the context of culture and language revitalization and assessment, see Kūkea Shultz & Englert, 2021, 2023). Some reviewers of this chapter argued that CLR principles may be more appropriate for classroom assessment than most large-scale testing programs; however, those engaged in the CLR work seek to influence testing across venues, purposes, and uses (Randall, 2021).

Some assessment researchers have promoted principles of culturally grounded validity evidence (what they called cultural validity) in the area of science assessments. Solano-Flores and Nelson-Barber (2001) argued that what they recommend would require a radical change in current test development and administration procedures, but the added effort may be necessary to achieve greater equity and fairness. They called into question the skills needed to develop fair assessments, such that test development teams should include cultural anthropologists and others who can bring a sociocultural perspective to the team. Specific disciplinary examples include mathematics in cultural contexts (Parker Webster & Yanez, 2007), mathematics in an authentic bicultural context (Lipka et al., 2007), and science

assessments acknowledging Aboriginal scientific knowledge and practices (Friesen & Ezeife, 2009).

A broad approach to this challenge is exemplified by the work of Canadian educators to incorporate Aboriginal content and epistemology throughout the education experiences of students (Claypool & Preston, 2011). An important fairness outcome of this work is the assertion that academic assessment must occur in a holistic context, recognizing physical, emotional, and spiritual forms of assessment. In a similar attempt to provide for fair testing of Aboriginal and Torres Strait Islander students in Australia, researchers are studying empirical data from large-scale testing, acknowledging the increasing cultural, ethnic, social, and linguistic diversity of the test-taker population. These researchers recommend greater balance of assessment types and alternative assessment approaches that are inclusive and participatory, particularly in the context of classroom assessment (Klenowski, 2016).

Another approach resulted in the transformation of large-scale assessments, including the Hawaiian state assessment program, supported by statewide collaboratives addressing Indigenous assessment. “Hawaiian culture-based education is based on a holistic view of the world and deep appreciation of interconnectedness” (Nā Lau Lama, 2007, chap. 2, p. 39), acknowledging that relevant knowledge and ways of knowing consider connectedness of worldviews and the assertion that “assessment within a Hawaiian context inherently includes the dimension of spirituality” (p. 38). Two principles of Hawaiian culture-based assessment include

E kuahui like i ka hana (Let everybody pitch in and work together): Assessment is strengths-based, respectful and constructive, looking for the particular attributes, contributions and potentials of the individuals or groups assessed, with particular emphasis on how they contribute to the larger community. Implicit in this approach is a respect for the “funds of knowledge” of students and their communities, an emphasis on growth, and continuous improvement that results from diligent effort. (p. 40)

Ma ka hana ka 'ike (In working, one learns): Assessment is personal in that it is appropriate to a particular individual, place and time. There is an emphasis on engagement as well as application of knowledge and skills in authentic ways. (p. 41)

Principles such as these directly impact the conceptualization of constructs and domains. These efforts extend what Moll and colleagues (1992) called funds of knowledge, the “historically accumulated and culturally developed bodies of knowledge and skills essential for household or individual functioning and well-being” (p. 133).

Test and Item Specifications in Support of CLR

Large-scale testing programs have item and task specifications describing sensitivity and bias concerns. Such specifications describe item content that should be avoided because of issues related to representation of diversity in content, illustrations, and contexts, with a focus on elements of stereotyping that should be avoided. Item

specifications also declare elements of language to avoid, including controversial or emotionally charged topics. In most cases, this includes avoidance of culturally embedded traditions, holidays, current events, politics, religion, and many other topics that might otherwise bring the test to life and make it relevant. Such sensitivity and bias-free practices not only leave the test context neutral (resolving to White, middle-class contexts), but also thus render it void of relevance to many test takers.

Haladyna and Rodriguez (2013) provided guidance for comprehensive item specifications, which also employ aspects of UDA and are supportive of the goals and components of PAD, beginning with precise descriptions of the construct, domain, knowledge and skills to be assessed, and specification of intended claims and relevant evidence. Item specifications also describe allowable item formats and their structures. Most important, item specifications describe the allowable sources and characteristics of reading passages, characteristics of stimulus materials (illustrations, figures, graphs), and characteristics of practice-based cases, scenarios, and vignettes. Item specifications lose ground regarding CLR in the descriptions (often long lists) of what to avoid, particularly regarding issues related to diversity; local, regional, or global affairs; and most topics that make school subjects, work, and life interesting.

The process of item development requires collaboration among a number of individuals and groups, including general test developers, item writers who typically are content or subject matter experts, measurement specialists (who may be psychometricians) and relevant specialists in areas such as culture, language development, gender issues and cognitive, emotional/behavioral or physical disabilities. (Rodriguez, 2016, p. 263)

Another potentially powerful contributor to item development is the test taker, where critical insight could be obtained through including student voices in K–12 education assessment development (Roach & Beddow, 2011), as well as student and community voices more generally. This collaborative process is rarely fully realized because item development and review typically occur sequentially, sometimes iteratively, but rarely collaboratively, with multiple specialists reviewing items, just not together. The *Standards* contains many instances of standards articulated with conditions such as “if appropriate” or “if possible.” For example, Standard 4.6 states,

When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. (AERA et al., 2014, p. 87)

We argue that such external review would be appropriate in all cases, particularly in the early stages of test specification development.

Assessment for Justice

There exists another movement in the educational testing and assessment arena that pushes CLR even further. It provides for a grander purpose than responsiveness. Some

have long argued that testing and assessment can serve an educational purpose, that individuals learn preparing for the test, taking the test, and reviewing the test (Mehrens & Lehman, 1991). There exist strong examples of tests of, for, and as learning (Bennett, 2015). To push these ideas in a progressive direction, educators are promoting the use of assessment to achieve justice goals—following the direction of efforts to promote social justice-oriented instruction (McLaren, 2000; Papa et al., 2016; Williamson et al., 2007). The introduction of social justice theory and goals in instruction and assessment has as its goal advancing opportunity, particularly for students who have been historically disenfranchised.

In the assessment realm, researchers have illuminated the important connections between assessment, justice, and opportunity (Gardner et al., 2009; McArthur, 2016). In the case of writing assessment, researchers, in collaboration with educators, ask the question, “How can we ensure that writing assessment leads to the advancement of opportunity?” (Poe et al., 2018, p. 379). They argue, and demonstrate, that by disrupting the historic disciplinary isolation of writing, writing assessment can promote student agency and advance opportunities for all students, consistent with the justice orientation of the field of writing and English education. This includes a reorientation of instructional approaches to writing, eliminating barriers to the potential source material and construction of writing exercises, and giving authentic purposes to writing and writing assessments, essentially unconstraining construct representation in writing assessment “to produce evidence related to fairness of comparable consequences for all” (Poe et al., 2018, p. 15).

In a state high school test of writing skills during the early 2000s, the following seemingly neutral and uncontroversial prompt was given to students: If you could change one thing about your life, what would it be and why? Unexpectedly, many students wrote stories about horrific neglect and abuse and illegal behaviors, both as victims and as perpetrators. Readers scoring the essays began to ask whether they were considered mandatory reporters and whether they were required to report cases to authorities—the identities of the students were known because the essay was part of the state accountability testing program. Were these truthful stories or were students writing fiction to display their skills? In addition, the state school counseling association took issue with the potential negative focus of the question because their collective mission was one of positive youth development with a focus on positive assets and opportunities for students. A seemingly neutral writing prompt required mass destruction of the written responses and the provision of counseling for test-response debriefing supports.

In the context of writing, justice prompts abound. Gonchar (2018) offered over 1,000 writing prompts, followed by 300 additional argument writing prompts. Among them were the following:

Should kids be social media influencers? Should all Americans receive anti-bias education? Should all companies require anti-bias training for employees? Is it offensive for sports teams and their fans to use native American names, imagery, and gestures? Do you support affirmative action in college admissions? What

rules should apply to transgender athletes when they compete? Why is race so hard to talk about?

Such prompts provide space to display writing skill, argumentation skills, and opportunity to express voice and to do so in a context that allows for multiple perspectives. Even among the GRE analytical writing prompts we find opportunities for students to explore issues of justice, possibly not because ETS is intentionally adopting issues of justice as part of a progressive mission, but because such issues are relevant to adults globally and acknowledge that there are multiple views, experiences, and perspectives. For example, one issue prompt the test employs asks test takers to discuss the statement: “The well-being of a society is enhanced when many of its people question authority” (https://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/pool). An example of an argument prompt is based on a scenario of an anthropologist studying childrearing in an island society that identifies two very different approaches to childrearing when comparing observation-centered versus interview-centered methodologies—calling into question the validity of the interpretations of one set of results (https://www.ets.org/gre/revised_general/prepare/analytical_writing/argument/pool). In both examples, individuals who live in or experience multiple cultural settings will find the prompts relevant and be able to express their voice on issues of justice, perhaps motivating relevant engagement in the measurement procedure.

Although there is not a strong presence in the assessment literature, some are introducing practices targeting antiracist education practices, including instruction and assessment that promote the recognition and analysis of injustice and introduce disciplinary approaches to change (Inoue, 2015, 2019; Kishimoto, 2018). Through antiracist practices in teaching, learning, and assessment, assessment tasks are developed to disrupt racist beliefs in disciplinary knowledge and practices by directly confronting economic, structural, and historical roots of inequality (McGregor, 1993; although McGregor wrote about this in the context of teaching). In the assessment arena, antiracist approaches to assessment ensure that tasks are developed to sustain test-taker cultural and linguistic disciplinary knowledge and practices, rather than ignore or eradicate those ways of knowing and doing (Baker-Bell, 2020). Randall et al. (2023) introduced a justice-oriented antiracist framework for validation. They applied this framework to address construct articulation, data analysis, and score interpretation, uncovering racist logics within standard processes of assessment design. Randall (2023) provided a justice-oriented antiracist perspective on bias and sensitivity review, processes that occur at multiple stages of item development and review, including, for example, following DIF analyses.

Russell (2024) provided a comprehensive review of the role of oppression and racism, as well as White supremacy, to move the field toward antiracist approaches and increase fairness. Russell demonstrated how the White racial frame structures test development theories and practices in ways that perpetuate racialized social structures.

Although these scholars are leading the way, test development, administration, analysis, and use practices remain grounded in prior limited notions and applications of fairness.

Implications for Practice and Research

Test developers need to manage the somewhat conflicting guidance in the *Standards* by revising long-standing item-writing guidelines and sensitivity review procedures to focus not only on what to avoid in item and test content and contexts, but also what to include to embrace CLR. Researchers should develop agendas to evaluate the impact of such approaches. A CLR assessment research agenda should include the application of UDL guidelines that embrace the principles of multiple means in terms of the why, what, and how of learning and doing. Agencies, professional associations, and other entities charged with defining content domains and content standards regarding what people are expected to know and do must engage their constituencies in discussions of essential, necessary, and CLR content. To promote fair tests, from the purpose of testing, test domain specifications, item-writing guidelines, test construction methods, field-test procedures, and operational administration and scoring, the roles of cultural and linguistic disciplinary content knowledge and practices should be considered. We must begin to ask questions about the oppressive and racist frameworks under which we have worked in the past to promote antiracist and fair approaches for a better future to promote social equity, access, and opportunity; testing and assessment more broadly can be an important tool to achieve greater justice.

CONCLUDING REMARKS

Fairness in educational measurement is promoted through consistent commitment to inclusion, access, and rigorous attention to content coverage and relevance and the elimination of barriers to access and sources of construct-irrelevant variance. There are many conditions under which fair and valid score interpretation and use are supported and enhanced, beginning with the way decisions are made about the need and role for testing. A preliminary condition for fair test interpretation and use requires a comprehensive analysis of opportunity to learn, because unfair, discriminatory, or exclusionary practices prohibit fair and uniform score interpretation (AERA et al., 2014, see p. 54). This includes the biases employed in the definition and specification of constructs and content domains, as well as every phase of test development, administration, scoring, reporting, and use. For some test takers, fairness also depends on the use of accommodations because even the most accessible test may not be accessible to all. In addition, fairness regarding access to the construct or domain is about what is *not* included in the content domain as much as it is about what *is*.

“The National Council on Measurement in Education is a community of measurement scientists and practitioners who work together to advance theory and applications

of educational measurement to benefit society” (<https://www.ncme.org/about/mission>). In his NCME presidential address, Sireci (2021) proposed five values, based on scholars and theorists in the field, and consistent with the *Standards*, including the following:

1. Everyone is capable of learning.
2. There are no differences in the capacity to learn across groups defined by race, ethnicity, or sex.
3. All educational tests are fallible to some degree.
4. Educational tests can provide valuable information to (a) improve student learning, and (b) certify competence.
5. All uses of educational test scores must be sufficiently justified by validity evidence. (p. 12)

These are fundamental values that undergird fairness in testing. The NCME mission statement calls on us to “serve the common good” (Sireci, 2021), a focus that was elevated during Sireci’s leadership of the organization. To promote that goal, he called on the organization to

- (a) ensure we enforce adherence to our AERA et al. (2014) *Standards*, particularly as they relate to the provision of validity evidence to defend test use;
- (b) de-emphasize norm-referenced competitiveness in educational testing except in those rare instances where examinees actually are competing for a benefit;
- (c) reorient our practices so that we value students more than the score scale;
- (d) engage with teachers and other educators to collaboratively develop tests and interpret test scores;
- (e) reconceptualize our notions of standardization to make tests more flexible to students’ needs and funds of knowledge;
- (f) design test score reports for students that emphasize their strengths, rather than their weaknesses; and
- (g) take full advantage of technology to allow assessments to tailor themselves to the needs of each specific examinee, foster engagement in the testing process, and to be fully aligned with and integrated into instruction. (p. 14)

These steps offer another way to engage in the principles of accessibility.

Relying on the *Standards* (AERA et al., 2014) and the principles of fairness discussed throughout the chapter, a number of concerns and guidance include broader attention to balanced assessment and policy design (Brookhart et al., 2019; Chattergoon & Marion, 2016; Martineau et al., 2018). The primary focus of balanced assessment efforts requires a careful articulation of the purposes of assessment, followed by the identification or creation and selection of assessments to meet those purposes, where the more important and valued purposes encompass the majority of assessment efforts, reducing emphasis on less essential or informative assessment practices. In addition, no decision resulting in consequences for test takers, groups, or organizations should be made based on a single test score, especially without retake opportunities. A long-standing principle of good measurement practice is to employ multiple measures, which underlies much of the discussion throughout this chapter. Multiple measures capture a richer picture of the intended construct or content knowledge and skills and enhance our understanding of what people know and can do. Consistent

with fairness concerns, multiple measures may help to balance intraindividual measurement errors across assessments and allow for multiple ways for test takers to display their knowledge and skills.

Research-informed policies have strong grounding and promote fairness. As such, testing policies must be research informed to provide fairness for test takers from culturally and linguistically diverse backgrounds, as well as those with disabilities. Testing policies and their application will support fairness when they are research based, including policies regarding identification of individuals as English learners and those with disabilities.

A comprehensive test evaluation plan will include a comprehensive review of fairness in every step of the test development and operational stages. Additional focus needs to be given to the evaluation of item and test score functioning vis-à-vis cultural and linguistic characteristics, for example, evaluation of measurement invariance, differential item and distractor functioning, evidence from response processes, differential predictive validity evidence, and the search for unintended consequences among test-taker groups, acknowledging within-group heterogeneity.

A much larger constituency must continue to learn about the principles of testing and test interpretation and use in diverse contexts, particularly test takers with disabilities and English learners who may not be proficient in the test language, including test takers themselves, educators, families, community members, policy makers, teacher educators, workforce development and human resource specialists, and test developers. Test developers and users benefit greatly from reaching out to and engaging with members of diverse communities where tests will be administered and used.

Applications of fairness principles in score interpretation and use continue through the evaluation of potential intended and unintended consequences of testing. Of particular concern in the arena of test fairness is the potential reproduction of disparities and limited opportunities, which are also a function of policies outside the arena of education and employment, including housing, transportation, health, economic development, law enforcement and corrections, and others. The interpretation and use of educational and credentialing tests are not isolated, and public policies interact in unique and unintended ways. The role of consequences continues to be a debated issue (see Ercikan & Solano-Flores, Lane & Marion, and Zwick, all this volume, for additional discussions). Although we agree that the underlying disparities are the result of much larger and historical trends, policies, and social structures, our concern is about the extent to which tests may contribute to the maintenance of social stratification.

Principles of fairness are known and are being implemented and used in many settings. Educational measurement specialists and test developers and users have available to them a wide range of tools and resources to promote fairness in testing through accessibility for specific groups who have experienced challenges in test accessibility and, ultimately, for all test takers.

ACKNOWLEDGMENTS

We are grateful for the attention to detail and wise advice from reviewers throughout the process, including Michael Kane, Maria Araceli Ruiz-Primo, and Michael Russell, a special review by Julio Caesar, and the incredible support and patience of the editors, Linda Cook and Mary Pitoniak.

REFERENCES

- Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, 17(4), 371–392.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74(1), 1–28.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, 19(3), 16–26.
- Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc interpretations. *Teaching English Language*, 4(1), 1–18.
- Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice. In S.N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices* (pp. 181–198). Springer.
- Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. An NCME instructional module. *Educational Measurement: Issues and Practice*, 26(1), 36–46.
- Allalouf, A. (2011). *ITC guidelines for quality control in scoring, test analysis, and reporting of test scores*. International Test Commission.
- Allalouf, A. (2017). Quality control for scoring tests administered in continuous mode: An NCME instructional module. *Educational Measurement: Issues and Practice*, 36(1), 58–68.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. <https://www.testingstandards.net/open-access-files.html>
- August, D., & Hakuta, K. (1994). *Evaluating the inclusion of L.E.P. students in systemic reform*. U.S. Department of Education, Office of the Under Secretary, Planning and Evaluation Service. <https://web.stanford.edu/~hakuta/Publications>
- August, D., & Hakuta, K. (Eds). (1997). *Improving schooling for language minority children: A research agenda*. National Academies Press.
- Baker-Bell, A. (2020). We been knowin: Toward an antiracist language & literacy education. *Journal of Language & Literacy Education*, 16(1), 1–12. <https://files.eric.ed.gov/fulltext/EJ1253929.pdf>
- Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education*, 19(2), 115–132.

- Bauer, G. R., Churchill, S. M., Mahendran, M., Walwyn, C., Lizotte, D., & Villa-Rueda, A. A. (2021). Intersectionality in quantitative research: A systematic review of its emergence and applications of theory and methods. *SSM—Population Health*, 14, 1–11. <https://doi.org/10.1016/j.ssmph.2021.100798>
- Beddow, P. A. (2012). Accessibility theory for enhancing the validity of test results for students with special needs. *International Journal of Disability, Development and Education*, 59(1), 97–111.
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370–407.
- Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., & Chan, D. (2015). *School composition and the Black–White achievement gap* (NCES 2015-018). U.S. Department of Education, National Center for Education Statistics.
- Bolt, S., & Roach, A. T. (2009). *Inclusive assessment and accountability: A guide to accommodations for students with diverse needs*. Guilford Press.
- Brookhart, S., Stiggins, R., McTighe, J., & Wiliam, D. (2019). *The future of assessment practices: Comprehensive and balanced assessment systems*. Learning Sciences International.
- Buzick, H., & Stone, E. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement: Issues and Practice*, 33(3), 17–30.
- Caesar, J. (2020). *Standardized bilingual assessments: A means to reduce construct-irrelevant variance and ethnic/racial stereotype threat* [Doctoral dissertation]. University of Minnesota Digital Conservancy. <https://conservancy.umn.edu/handle/11299/216165>
- Cawthon, S., & Leppo, R. (2013). Assessment accommodations on tests of academic achievement for students who are deaf or hard of hearing: A qualitative meta analysis of the research literature. *American Annals of the Deaf*, 158(3), 363–376.
- Chattergoon, R., & Marion, S. (2016). Not as easy as it sounds: Designing a balanced assessment system. *State Education Standard*, 16(1), 6–9. <https://eric.ed.gov/?id=EJ1087555>
- Christensen, L. L., Shyyan, V. V., & Stewart, K. (2018). *Developing a request for proposals for an alternate assessment of English language proficiency*. Alternate English Language Assessment. <https://altella.wceruw.org/pubs/Developing-Request-for-Proposals.pdf>
- Claypool, T. R., & Preston, J. P. (2011). Redefining learning and assessment practices impacting Aboriginal students: Considering Aboriginal priorities via Aboriginal and western worldviews. *in education*, 17(3), 84–95.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Department of Health, Education, and Welfare, Office of Education. <https://files.eric.ed.gov/fulltext/ED012275.pdf>
- Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly*, 23(3), 509–531.
- Cottrell, J. M., Newman, D. A., & Roisman, G. I. (2015). Explaining the Black–White gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology*, 100(6), 1713–1736.

- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1), 139–167. <http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8>
- Dee, T. S., & Domingue, B. W. (2021). Assessing the impact of a test question: Evidence from the “Underground Railroad” controversy. *Educational Measurement: Issues and Practice*, 40(2), 81–88. <https://doi.org/10.1111/emip.12411>
- Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W., & Rose, D. H. (2013). *A universal design for learning-based framework for designing accessible technology-enhanced assessments*. Pearson. <https://eric.ed.gov/?id=ED576691>
- Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English language arts assessment of elementary, middle, and high school students*. Pearson Education. http://images.pearsonassessments.com/images/tmrs/cognitive_lab_evaluation_of_innovative_items.pdf
- Duran, L. K., Wackerle-Hollman, A., Kohlmeier, T. L., Brunner, S. K., Palma, J., & Callard, C. H. (2019). Individual Growth and Development Indicators–Español: Innovations in the development of Spanish oral language general outcome measures. *Early Childhood Research Quarterly*, 48, 155–172.
- ETS. (2022). *ETS guidelines for developing fair tests and communications*. <https://www.ets.org/content/dam/ets-org/pdfs/about/fair-tests-and-communications.pdf>
- Eichorn, L. (1997). Reasonable accommodations and awkward compromises: Issues concerning learning disabled students and professional schools in the law context. *Journal of Law and Education*, 26 (1), 31–63.
- Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (Eds.). (2018). *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (2nd ed.). Springer.
- Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 69–86). Emerald Group.
- Erickson, K. A., & Geist, L. A. (2016). The profiles of students with significant cognitive disabilities and complex communication needs. *Augmentative and Alternative Communication*, 32(3), 187–197.
- Erickson, K., & Quick, N. (2017). The profiles of students with significant cognitive disabilities and known hearing loss. *Journal of Deaf Studies and Deaf Education*, 22(1), 35–48.
- Faulkner-Bond, M. (2020). Comparability when assessing English learner students. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 150–175). National Academy of Education. <https://naeducation.org/wp-content/uploads/2020/06/Comparability-of-Large-Scale-Educational-Assessments.pdf>

- Frankenberg, E., Ee, J., Ayscue, J. B., & Orfield, G. (2019). *Harming our common future: America's segregated schools 65 years after Brown*. Center for Education and Civil Rights, University of California, Los Angeles. <https://www.civilrightsproject.ucla.edu>
- Friesen, J. B., & Ezeife, A. N. (2009). Making science assessment culturally valid for Aboriginal students. *Canadian Journal of Native Education*, 32(2), 24–37.
- Fuchs, L. S., Fuchs, D., Enton, S. B., Hamlett, C. L., & Karns, K. M. (2000). Supplementing teacher judgements of mathematics test accommodations with objective data sources. *School Psychology Review*, 29(1), 65–85.
- Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment*, 8(1), 27–41.
- Garcia, G. N. (2000). *Lessons from research: What is the length of time it takes limited English proficient students to acquire English and succeed in an all-English classroom?* National Clearing House for Bilingual Education, Issue Brief No. 5. <https://files.eric.ed.gov/fulltext/ED450585.pdf>
- Gardner, J., Holmes, B., & Leitch, R. (2009). *Assessment and social justice: A Futurelab literature review*. Futurelab. <https://www.nfer.ac.uk/assessment-and-social-justice/>
- Garrett Holiday, B. (2009). The history and visions of African American psychology: Multiple pathways to place, space, and authority. *Cultural Diversity and Ethnic Minority Psychology*, 15(4), 317–337.
- Geisinger, K. F. (Ed.). (2015). *Psychological testing of Hispanics: Clinical, cultural, and intellectual issues* (2nd ed). American Psychological Association.
- Gonchar, M. (2018, April 12). Over 1,000 writing prompts for students [Student opinion piece]. *The New York Times*. <https://www.nytimes.com/2018/04/12/learning/over-1000-writing-prompts-for-students.html>
- Gregg, N., & Nelson, J. (2012). Meta-analysis of the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, 45(2), 128–138.
- Hakuta, K., Goto Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* The University of California Linguistic Minority Research Institute, Policy Report 2000-1. <https://escholarship.org/uc/item/13w7m06g>
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.
- Heaney, K. J., & Pullin, D. C. (1998). Accommodations and flags: Admissions testing and the rights of individuals with disabilities. *Educational Assessment*, 5(2), 71–93.
- Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. WAC Clearinghouse. <https://open.umn.edu/opentextbooks/textbooks/293>
- Inoue, A. B. (2019). Classroom writing assessment as an antiracist practice: Confronting White supremacy in the judgments of language. *Pedagogy*, 19(3), 373–404.

- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15.
- Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.
- Kanno, Y., & Cromley, J. G. (2013). English language learners' access to and attainment in postsecondary education. *TESOL Quarterly*, 47(1), 89–121.
- Karami, H., & Salmani Nodoushan, M. A. (2011). Differential item functioning (DIF): Current problems and future directions. *International Journal of Language Studies*, 5(3), 133–142.
- Kearns, J. F., Towles-Reeves, E., Kleinert, H. L., Kleinert, J. O., & Thomas, M. K.-K. (2011). Characteristics of and implications for students participating in alternate assessment based on alternate academic achievement standards. *The Journal of Special Education*, 45(1), 3–14.
- Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practices*, 27(3), 3–16.
- Ketterlin-Geller, L. R., Johnstone, C. J., & Thurlow, M. L. (2015). Universal design in assessment. In S. E. Burgstahler (Ed.), *Universal design in higher education: From principles to practice* (2nd ed., pp. 163–175). Harvard Education Press.
- Kettler, R. J., Elliott, S. N., & Beddow, P. A. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education*, 84, 529–551.
- Kettler, R. J., Elliott, S. N., Beddow, P. A., & Kurz, A. (2018). Accessible instruction and testing today. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and applications* (pp. 1–16). Springer International Publishing. <https://doi.org/10.1007/978-3-319-71126-3>
- Kieffer, M. J., Lesaux, N. K., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, 79(3), 1168–1201.
- Kishimoto, K. (2018). Anti-racist pedagogy: From faculty's self-reflection to organizing within and beyond the classroom. *Race Ethnicity and Education*, 21(4), 540–554.
- Klenowski, V. (2016). Fairer assessment for Indigenous students: An Australian perspective. In S. Scott, D. Scott, & C. Webber (Eds.), *Leadership of assessment, inclusion, and learning* (pp. 273–285). Springer International.
- Knoester, M., & Au, W. (2015). Standardized testing and school segregation: Like tinder for fire? *Race Ethnicity and Education*, 20(1), 1–14.
- Kopriva, R., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist*, 5(1), 108–128.
- Kosak-Babuder, M., Kormos, J., Ratajczak, M., & Pizorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing*, 36(1), 51–75.

- Kūkea Shultz, P., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education*, 6, 1–11.
- Kūkea Shultz, P., & Englert, K. (2023). The promise of assessments that advance social justice: An Indigenous example. *Applied Measurement in Education*, 36(3), 255–268.
- LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review*, 64(1), 55–75.
- Laing, J., & Farmer, M. (1984). *Use of ACT assessment by examinees with disabilities* (Research Report No. 84). ACT. <https://eric.ed.gov/?id=ED249261>
- Laitusis, C. C., & Cook, L. L. (2007). *Large-scale assessment and accommodations: What works?* (pp. 67–79). Council for Exceptional Children.
- Larson, E. D., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2020). Paradigm shifts in states' assessment accessibility policies: Addressing challenges in implementation. *Journal of Disability Policy Studies*, 30(4), 244–252. <https://doi.org/10.1177/1044207319848071>
- Lawton, M. (1995, September 6). NAEP to cut exams, test more disabled pupils. *Education Week*, 27.
- Lazarus, S., Goldstone, L., Wheeler, T., Paul, J., Prestridge, S., Sharp, T., Hochstetter, A., & Warren, S. (2021). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students*. Council of Chief State School Officers.
- Lazarus, S. S., & Thurlow, M. L. (2016). *2015–16 high school assessment accommodations policies: An analysis of ACT, SAT, PARCC, and Smarter Balanced* (NCEO Report 403). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/Report403/NCEOREport403.pdf>
- Lazarus, S. S., Thurlow, M. L., Lail, K. E., & Christensen, L. (2009). A longitudinal analysis of state accommodations policies: Twelve years of change 1993–2005. *Journal of Special Education*, 43(2), 67–80.
- Lazarus, S. S., Thurlow, M. L., Ysseldyke, J. E., & Edwards, L. M. (2015). An analysis of the rise and fall of the AA-MAS policy. *Journal of Special Education*, 48(4), 231–242.
- Lazarus, S. S., van den Heuvel, J. R., & Thurlow, M. L. (2017). Using K–12 lessons learned about how to balance accessibility and test security to inform licensure, credentialing, and certification exam policies. *Journal of Applied Testing Technology (JATT)*, 18(1), 1–11.
- Li, H. (2014). The effects of read-aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement: Issues and Practice*, 33(3), 3–16.
- Linquanti, R., Cook, H. G., Bailey, A. L., & McDonald, R. (2013). *Toward a more common definition of English learners: Collected guidance for states and multi-state assessment consortia*. Council of Chief State School Officers.
- Lipka, J., Sharp, N., Adams, B., & Sharp, F. (2007). Creating a third space for authentic biculturalism: Examples from math in a cultural context. *Journal of American Indian Education*, 46(3), 94–115.

- Little, D. (1999). Learning disabilities, medical students, and common sense. *Academic Medicine*, 74, 622–623.
- Liu, K. K., Thurlow, M. L., Press, A. M., & Lickteig, O. (2018). *A review of the literature on measuring English language proficiency progress of English learners with disabilities and English learners* (NCEO Report 408). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/NCEOReport408.pdf>
- Martineau, J., Dewsbury-White, K., Roeber, E., Vorenkamp, E., Snead, S., & Flukes, J. (2018). *District assessment system design toolkit*. Center for Assessment. <https://www.nciea.org/featured-resources>
- McArthur, J. (2016). Assessment for social justice: The role of assessment in achieving social justice. *Assessment & Evaluation in Higher Education*, 41(7), 967–981.
- McDonnell, L. M., & McLaughlin, M. J. (1997). *Educating one & all: Students with disabilities and standards-based reform*. National Academies Press. https://nap.nationalacademies.org/login.php?record_id=5788
- McGregor, J. (1993). Effectiveness of role playing and antiracist teaching in reducing student prejudice. *Journal of Educational Research*, 86(4), 215–226.
- McLaren, P. (2000). *Che Guevara, Paulo Freire, and the pedagogy of revolution*. Rowman & Littlefield.
- Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Wadsworth.
- Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist*, 20, 136–142.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–1027.
- Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10, 9–20.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25, 6–20. <https://doi.org/10.1111/j.1745-3992.2006.00075.x>
- Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice*, 31(2), 132–141.
- Moore, J., Huang, C., Nooree, H., Li, T., & Camara, W. (2018). *Testing supports for English learners: A literature review and preliminary ACT research findings* (Working Paper 2018-1). ACT. <https://files.eric.ed.gov/fulltext/ED593176.pdf>
- Nā Lau Lama. (2007). *Nā Lau Lama community report*. Kamehameha Schools. http://www.ksbe.edu/_assets/spi/pdfs/reports/na-lau-lama/Executive_Summary_Final.pdf

- National Center for Education Statistics. (1988). *National Education Longitudinal Study of 1988 (NELS:88)*. U.S. Department of Education. <https://nces.ed.gov/surveys/nels88/>
- National Center for Education Statistics. (2019). *Status and trends in the education of racial and ethnic groups (as of February 2019)*. U.S. Department of Education. <https://nces.ed.gov/programs/raceindicators/index.asp>
- National Center for Education Statistics. (2021). *Percentage distribution of public school students, for each racial and ethnic group, by school poverty level: Fall 2017*. U.S. Department of Education. <https://nces.ed.gov/fastfacts/display.asp?id=898>
- National Center for Education Statistics. (2023). *NAEP accommodations increase inclusiveness*. https://nces.ed.gov/nationsreportcard/about/accom_table.aspx
- National Center on Educational Outcomes. (2014). *Exploring alternate ELP assessments for ELLs with significant cognitive disabilities* (NCEO Brief 10). University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/briefs/brief10/NCEOBrief10.pdf>
- National Center on Educational Outcomes. (2020). *English learners (ELs)*. University of Minnesota. https://nceo.info/student_groups/english_language_learners
- National Council on Disability. (2018). *English learners and students from low income families*.
- National Council on Measurement in Education. (2020). *Position statement on testing English learners*. <https://www.ncme.org/resources-publications/position-statements/equity>
- Nester, M. A. (1993). Psychometric testing and reasonable accommodations for persons with disabilities. *Rehabilitation Psychology*, 38(2), 75–85.
- Oliveri, M. E., & von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing*, 16, 220–239.
- Olson, J. F., & Goldstein, A. A. (1996). Increasing the inclusion of students with disabilities and limited English proficient students in NAEP. *Focus on NAEP*, 2(1) 1–5. <https://files.eric.ed.gov/fulltext/ED400339.pdf>
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Lawrence Erlbaum Associates.
- O'Neill, K. A., McPeck, W. M., & Wild, C. L. (1993). *Differential item functioning on the Graduate Management Admissions Test* (ETS Research Report No. RR-93-35). ETS. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01546.x>
- Papa, R., Eadens, D. M., & Eadens, D. W. (Eds.). (2016). *Social justice instruction: Empowerment on the chalkboard*. Springer.
- Parker Webster, J., & Yanez, E. (2007). Qanemcikarluni tekitnarqelartuq [One must arrive with a story to tell]: Traditional Alaska native Yup'ik Eskimo stories in a culturally based math curriculum. *Journal of American Indian Education*, 46(3), 116–131.

- Penuel, W. R., & Shepard, L. A. (2016). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). John Wiley. <https://doi.org/10.1002/9781118956588.ch7>
- Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education*, 7(2), 93–120.
- Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, 71(1), 53–104. <https://doi.org/10.3102/00346543071001053>
- Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. University Press of Colorado.
- Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 89(2), 146–150. http://www.pdkmembers.org/members_online/publications/Archive/pdf/k0710pop.pdf
- Popham, W. J., & Ryan, J. M. (2012, April 12–16). *Determining a high-stakes test's instructional sensitivity* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Vancouver, BC, Canada.
- Quenemoen, R. F. (2009). The long and winding road of alternate assessments: Where we started, where we are now, and the road ahead. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 127–153). Paul H. Brookes.
- Race to the Top Fund, (34 CFR Subtitle B, Chapter II) 74 Fed. Reg. 59687. (November 18, 2009). <https://www.govinfo.gov/content/pkg/FR-2009-11-18/pdf/E9-27426.pdf>
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 32–34.
- Randall, J. (2023). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment*, 28(2), 68–82.
- Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2023). Justice-oriented, antiracist validation: Continuing to disrupt White supremacy in assessment practices. *Educational Assessment*, 29 (1), 1–20. <https://doi.org/10.1080/10627197.2023.2285047>
- Reardon, S. F., Weathers, E. S., Fahle, E. M., Jang, H., & Kalogrides, D. (2019). *Is separate still unequal? New evidence on school segregation and racial academic achievement gaps* (CEPA Working Paper No.19-06). Stanford Center for Education Policy Analysis, Stanford University. <http://cepa.stanford.edu/wp19-06>
- Rios, J. A., Ihlenfeldt, S. D., & Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multistudy systematic review and meta-analysis. *Educational Measurement: Issues and Practice*, 39(4), 65–75.

- Roach, A. T., & Beddow, P. A. (2011). Including student voices in the design of more inclusive assessments. In S. Elliott, R. Kettler, P. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students* (pp. 243–254). Springer.
- Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Lawrence Erlbaum Associates.
- Rodriguez, M. C. (2009). Psychometric considerations for alternate assessments based on modified academic achievement standards. *Peabody Journal of Education*, 84(4), 595–602.
- Rodriguez, M. C. (2011). Item-writing practice and evidence. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy* (pp. 201–216). Springer.
- Rodriguez, M. C. (2016). Selected-response item development. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 259–273). Routledge.
- Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *Sage Open*, 4(4), 1–10.
- Rodriguez, M. C., & Nickodem, K. (2018, April 12–16). *Comprehensive partitioning of student achievement variance to inform equitable policy design* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States. <https://conservancy.umn.edu/handle/11299/195229>
- Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2014). *A summary of the research on the effects of test accommodations, 2011–2012* (Synthesis Report 94). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/onlinepubs/Synthesis94/Synthesis94.pdf>
- Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2016). *A summary of the research on the effects of test accommodations: 2013–2014* (NCEO Report 402). National Center on Educational Outcomes, University of Minnesota. <https://nceo.info/Resources/publications/OnlinePubs/Report402/>
- Rogers, C. M., Lazarus, S. S., Thurlow, M. L., & Liu, K. K. (2020). *A summary of the research on the effects of K–12 test accommodations: 2017* (NCEO Report 418). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/NCEOReport418.pdf>
- Rogers, C. M., Thurlow, M. L., Lazarus, S. S., & Liu, K. K. (2019). *A summary of the research on effects of test accommodations: 2015–2016* (NCEO Report 412). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/NCEOReport412.pdf>
- Rose, D. H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies*. Harvard Education Press.
- Rubel, L., & McCloskey, A. V. (2019). The “soft bigotry of low expectations” and its role in maintaining White supremacy through mathematics education. *Critical Mathematical Inquiry*, 41, 113–128. <https://educate.bankstreet.edu/occasional-paper-series/vol2019/iss41/10>

- Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, 17(1), 20–32. <http://www.jattjournal.com/index.php/atp/article/view/89189/67798>
- Russell, M. (2022). Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement: Issues and Practice*, 41(2), 25–35.
- Russell, M. (2024). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond*. Routledge.
- Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K–12 testing programs. *Educational Assessment*, 24(4), 286–304.
- Russell, M., Szendey, O., & Li, Z. (2022). An intersectional approach to DIF: Comparing outcomes across methods. *Educational Assessment*, 27(2), 115–135.
- Schaeffer, K. (2020). *6 facts about economic inequality in the U.S.* Pew Research Center. <https://www.pewresearch.org/fact-tank/2020/02/07/6-facts-about-economic-inequality-in-the-u-s/>
- Schifter, L. A., Grindal, T., Schwartz, G., & Hehir, T. (2019). *Students from low-income families and special education*. The Century Foundation. <https://tcf.org/content/report/students-low-income-families-special-education/>
- Semega, J., Kollar, M., Creamer, J., & Moharitty, A. (2020). *Income and poverty in the United States: 2018, Current population reports*. U.S. Census Bureau. <https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-266.pdf>
- Shaftel, J., Benz, S., Boeth, E., Gahm, J., He, D., Loughran, J., Mellen, M., Meyer, E., Minor, E., & Overland, E. (2015). *Accessibility for technology-enhanced assessments: Report of project activities*. Center for Educational Testing and Evaluation, University of Kansas. <https://ateassessments.atlas4learning.org/documents>
- Shanedling, J., Van Heest, A., Rodriguez, M. C., Putnam, M., & Agel, J. (2010). Validation of an online assessment of orthopedic surgery residents' cognitive skills and preparedness for carpal tunnel release surgery. *Journal of Graduate Medical Education*, 2(3), 435–441.
- Sharkey, P., & Elwert, F. (2012). The legacy of disadvantage multigenerational neighborhood effects on cognitive ability. *American Journal of Sociology*, 116(6), 1934–1981.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018a). Using learning and motivation theories to coherently link formative assessment, grading practices, and large-scale assessment. *Educational Measurement: Issues and Practice*, 37(1), 21–34.
- Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018b). Classroom assessment principles to support learning and avoid the harms of testing. *Educational Measurement: Issues and Practice*, 37(1), 52–57.
- Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, 40(1), 7–16.

- Sireci, S. G., Scarpatti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490.
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
- Stone, E., Cahalan Laitusis, C., & Cook, L. L. (2015). Increasing the accessibility of assessments through technology. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 217–234). Routledge.
- Strain-Seymour, E., Way, W. D., & Dolan, R. P. (2009). *Strategies and processes for developing innovative items in large-scale assessments*. Pearson Education. <https://images.pearsonassessments.com/images/tmrs/StrategiesandProcessesforDeveloping-InnovativeItems.pdf>
- Stricker, L. J., & Emmerich, W. (1999). Possible determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement*, 36(4), 347–366.
- Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). Considerations for the development and review of universally designed assessments. National Center on Educational Outcomes, University of Minnesota. <https://rtc3.umn.edu/docs/OnlinePubs/TechReport42.pdf>
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). National Center on Educational Outcomes, University of Minnesota. <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>
- Thompson, S. J., Thurlow, M. L., & Malouf, D. (2004). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology*, 6(1), 1–15.
- Thurlow, M. L., Elliott, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements* (2nd ed). Corwin.
- Thurlow, M. L., Elliott, J. L., Erickson, R. N., & Ysseldyke, J. E. (1997). Learning disabilities and accommodations: Best practice for bar exams. *The Bar Examiner*, 66(4), 17–30.
- Thurlow, M. L., & Kopriva, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education*, 39, 331–369.
- Thurlow, M. L., Lazarus, S. S., & Bechard, S. (Eds). (2013). *Lessons learned in federally funded projects that can improve the instruction and assessment of low performing students with disabilities*. National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/LessonsLearned.pdf>
- Thurlow, M. L., Lazarus, S. S., Albus, D. A., Larson, E. D., & Liu, K. K. (2019). *2018–19 participation guidelines and definitions for alternate assessments based on alternate academic achievement standards* (NCEO Report 415). National Center on

- Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/NCEOReport415.pdf>
- Thurlow, M. L., Lazarus, S. S., Christensen, L. L., & Shyyan, V. (2016). *Principles and characteristics of inclusive assessment systems in a changing assessment landscape* (NCEO Report 400). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/Report400/NCEOReport400.pdf>
- Thurlow, M. L., Liu, K. K., Ward, J. M., & Christensen, L. L. (2013). *Assessment principles and guidelines for ELLs with disabilities. Improving the validity of assessment results for English language learners with disabilities (IVARED)*. National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/ivared/IVAREDPrinciplesReport.pdf>
- Thurlow, M. L., & Quenemoen, R. F. (2016). Alternate assessments for students with disabilities: Lessons learned from the National Center and State Collaborative. In C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 416–432). Guilford.
- Thurlow, M. L., Shyyan, V. V., Lazarus, S. S., & Christensen, L. L. (2017). *Providing English language development services to English learners with disabilities: Approaches to making exit decisions* (NCEO Report 404). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/NCEOreport404.pdf>
- Thurlow, M. L., & Wu, Y.-C. (2019). *2016–2017 APR snapshot #20: Students in special education assigned assessment accommodations*. National Center on Educational Outcomes, University of Minnesota. <https://ici.umn.edu/products/sLyOfa4fS6yH-ubiml21iQ>
- Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Report 4). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/archive/Synthesis/SynthesisReport004.pdf>
- Thurlow, M. L., Ysseldyke, J. E., & Silverstein, B. (1995). Testing accommodations for students with disabilities. *Remedial and Special Education*, 16 (5), 260–270.
- Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *The Journal of Special Education*, 42(4), 241–254.
- U.S. Census Bureau. (2023). *Anniversary of Americans With Disabilities Act: July 26, 2023*. Press Release No. CB23-FF.06. <https://www.census.gov/newsroom/facts-for-features/2023/disabilities-act.html>
- U.S. Department of Education. (2010). *Recent trends in mean scores and characteristics of test-takers on Praxis II licensure tests*. Office of Planning, Evaluation, and Policy Development, Policy and Program Studies Service. <https://www2.ed.gov/rschstat/eval/teaching/praxis-ii/report.pdf>
- U.S. Department of Education. (2015, August 21). Improving the academic achievement of the disadvantaged: Assistance to states for the education of children with disabilities. *Federal Register*, 80(162), 50773–50785.

- U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*.
- U.S. Department of Labor. (2021). *Persons with a disability: Labor force characteristics—2020* (USDOL-21-0316). Bureau of Labor Statistics. <https://www.bls.gov/news.release/pdf/disabl.pdf>
- U.S. Government Accountability Office. (2011). *Higher education and disability: Improved federal enforcement needed to better protect students' rights to testing accommodations* (GAO-12-40). <https://www.gao.gov/assets/590/587367.pdf>
- Vanchu-Orosco, M. (2012). *Meta-analysis of testing accommodations for students with disabilities: Implications for high-stakes testing* [Doctoral dissertation, University of Denver]. Digital Commons @DU. <https://digitalcommons.du.edu/etd/668/>
- van den Heuvel, J. R., Lazarus, S. S., & Thurlow, M. L. (2016, June 13). Are accessibility and exam security mutually exclusive aims? *Certification Magazine*. <http://certmag.com/accessibility-exam-security-mutually-exclusive-aims/>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Wackerle-Hollman, A., Durán, L., Brunner, S., Palma, J., Kohlmeier, T., & Rodriguez, M. C. (2019). Developing a measure of Spanish phonological awareness for preschool age children: Spanish Individual Growth and Development Indicators. *Educational Assessment*, 24(1), 33–56.
- Wackerle-Hollman, A., Duran, L. K., Rodriguez, M. C., Chavez, C., Miranda, A., & Medina Morales, N. (2022). Understanding how language of instruction impacts early literacy growth for Spanish speaking children. *School Psychology Review*, 51, 406–426.
- Wiener, D. (2005). *One state's story: Access and alignment to the GRADE-LEVEL content for students with significant cognitive disabilities* (Synthesis Report 57). National Center on Educational Outcomes, University of Minnesota. <https://nceo.umn.edu/docs/OnlinePubs/Synthesis57.pdf>
- Williamson, J. A., Rhodes, L., & Dunson, M. (2007). A selected history of social justice in education. *Review of Research in Education*, 31, 195–224.
- Willingham, W. W., & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Routledge.
- Willingham, W. W., Rogasta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (Eds.). (1988). *Testing handicapped people*. Allyn & Bacon.
- Winter, P. C., Karvonen, M., & Christensen, L. L. (2018). *Developing item templates for alternate assessments of English language proficiency. Alternate English Language Assessment (ALTELLA)*. Wisconsin Center for Education Research, University of Wisconsin–Madison. https://altella.wceruw.org/pubs/ALTELLA_Report_Item_083108.pdf
- Wolfe, E. W., & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education*, 14(1), 91–107.
- Wu, Y.-C., Thurlow, M. L., & Liu, K. K. (2021). *Understanding the characteristics of English learners with IEPs to meet their needs during state and districtwide assessments*.

National Center on Educational Outcomes, University of Minnesota. <https://files.eric.ed.gov/fulltext/ED616099.pdf>

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15(4), 337–362.

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). Routledge.

Part IV

Specific Applications of Assessment

Assessment to Inform Teaching and Learning

Susan M. Brookhart

Duquesne University (Emerita)

Charlie DePascale

Independent Consultant

In recent years, at least from the advent of the standards movement in the 1980s (Porter, 1993), expectations have risen that all assessments—not just those designed to be used formatively—will inform teaching and learning. This trend defines the overarching theme of this chapter: Educators, students and parents, and community members increasingly look to many different assessments for information about improving learning; that is, they look to assessments to serve a formative purpose. This chapter describes the variety of assessments that stakeholders press into service to inform teaching and learning, evaluates their use in this endeavor, and suggests what the future may bring as the arc of assessment continues to bend toward the formative.

CHAPTER PURPOSE AND SCOPE

Both *teaching* and *learning* have multiple definitions, depending on one's philosophical and psychological approach (Gitomer & Bell, 2016). This chapter focuses on assessment of what students know and can do in relation to learning goals that are taught and learned in school classrooms, with the intent of (a) informing teacher instructional planning and instructional moves and (b) informing students' thinking and understanding during classroom lessons, including the individual and social regulation of learning. Similarly, *assessment* has multiple definitions, with some emphasizing assessment instruments and others emphasizing assessment processes. This chapter reviews research about both assessment instruments and the processes in which they are used. As Bennett (2011, p. 7) noted, "process cannot somehow rescue unsuitable instrumentation, nor can instrumentation save an unsuitable process." Much of the research reviewed is from the United States, where teaching to learning goals is the common model of instruction; however, learning-oriented assessment is also a growing interest globally (Carless, 2007; Zeng et al., 2018). The chapter does not focus on the use of assessment information to inform curriculum revision, materials review, or policy adoption.

THESIS

Both educators and the public increasingly expect assessment to be formative and to be helpful to the learning of individuals. Since at least the 1990s, assessment users have been clamoring for more formative information, even from assessments not designed to provide such information, like grades or state accountability tests, as evidenced by educators trying to use annual accountability assessments for instructional change and by the generally unsuccessful attempts to add diagnostic scoring to large-scale tests. The logic is often some version of an argument that runs, in general, that if an assessment tells something about what has been learned, it also should have implications for what should be learned next. Growing interest in student learning theory, fairness, and social justice may be influencing this movement toward the formative as well (see Ercikan & Solano-Flores; Lane & Marion; Rodriguez & Thurlow; and Zwick, this volume). This chapter reviews a range of assessments with an emphasis on how they are being used to

inform teaching and learning, whether they were originally intended for formative use or not.

Three subthemes emerge in this press toward the formative. First, the nature of the information is broad. Depending on its source and purpose, assessment that informs teaching and learning can result in information that is qualitative (e.g., feedback comments) or quantitative (e.g., scores, grades), or sometimes both (e.g., a rubric level associated with a performance-level description). Information can be about individuals or groups: Assessment that serves a formative purpose can support students working through a curriculum, and it can also help support broader decisions about instruction. Second, both information quality—the validity or trustworthiness of the assessment information—and the quality of implementation in practice are key to the effectiveness of assessment in informing teaching and learning (Bennett, 2011). Third, teacher and student assessment literacy is key to the effective use of assessment to inform teaching and learning.

ORGANIZATION OF THE CHAPTER

The chapter is organized into five sections, the first of which is this introduction. The second section describes assessments of students' current learning status: those focused on the formative, including classroom formative assessment and interim assessment, and those focused on the summative, including grading and state summative assessment. Evidence shows that even assessments intended for summative purposes are used to inform teaching and learning. The third section describes methods of assessment that assess trajectories of learning through the learning process. It begins by describing how the assessments in this section differ from conventional assessments and then describes assessments focused on trajectories of learning: curriculum-based measurement, student learning objectives, and assessment based on learning progressions. In both the second and the third sections, assessments are described according to their purpose and uses, design, measurement considerations, implementation in practice, and their impact on learning. The chapter ends with a fourth section on future directions for assessment to inform teaching and learning and a fifth, concluding section.

ASSESSMENTS OF STUDENTS' CURRENT LEARNING STATUS

More and more, educators, parents, and community members seek to leverage any available assessment information to inform teaching and learning, whether from assessments originally focused on the formative or from those originally focused on the summative.

Assessments Focused on the Formative

Assessment that has as its primary purpose informing teaching and learning is called formative assessment, and it is typically distinguished from summative assessment, which has as its primary purpose certifying or reporting learning. Bloom et al. (1971)

moved formative evaluation, originally conceived at the curriculum level, into the classroom. They emphasized assessment to inform teachers' instruction. During the 1990s, several scholars demonstrated the importance of students in the evaluation process as well (Black & Wiliam, 1998; Natriello, 1987; T. J. Crooks, 1988).

No one definition of formative assessment has yet taken its place in the field as the only authoritative one. However, definitions do converge on formative assessment as an evidentiary process, the need for student involvement, and the intent to move learning forward—just collecting information about student learning is not sufficient. Current definitions of formative assessment can be grouped into four nested categories, according to whether, to be called formative assessment, a process need only provide information about student thinking; must provide information that is potentially useful; must provide information that is used with the intention of improving student performance; or must provide information that is used and that has a positive effect on student performance (Furtak et al., 2015). The 2008 Council of Chief State School Officers' definition of formative assessment in practice (CCSSO, 2018) is used in the glossary of the *Standards for Educational and Psychological Testing* (the *Standards*; American Educational Research Association [AERA] et al., 2014, p. 219):

Formative assessment: An assessment process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning with the goal of improving students' achievement of intended instructional outcomes.

However, beyond the glossary entry, the *Standards* do not treat this kind of formative assessment. Other definitions have been noted (e.g., Andrade et al., 2019; CCSSO, 2018).

The term *assessment for learning* is very often used as a synonym for formative assessment. Often, whether one uses assessment for learning or *formative assessment* reflects one's research and practice traditions and cultural background. Some (e.g., Black et al., 2003; James & Pedder, 2006) use assessment for learning as a broad term and reserve use of the term formative assessment for instances when assessment information is actually used to improve learning. Conversely, Stiggins (2005) considered assessment for learning as one of several approaches to formative assessment, in which students are actively involved and "inside the assessment process, watching themselves grow, feeling in control of their success, and believing that continued success is within reach if they keep trying" (pp. 327–328).

For purposes of this chapter, Black and Wiliam's (2009, p. 9) definition of formative assessment will be used:

Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited.

Wiliam (2010, pp. 24–25) identified five key points in this definition:

1. “Anyone can be the agent in formative assessment,” teachers, learners, or peers.
2. “The focus of the definition is on decisions,” not intentions.
3. The focus is on “next steps in instruction,” which he defined as “any planful activity intended to create learning” (p. 25), whether by teacher or learner.
4. “The definition is probabilistic” (“likely to be better”) and acknowledges that not every well-founded decision will result in improved learning.
5. “The assessment need not change the planned instruction.” Sometimes, formative assessment evidence affirms the teacher or student’s original instructional plan.

This section describes three kinds of assessments that focus on formative uses. Informing teaching and learning is the main intended purpose of classroom formative assessment (Black & Wiliam, 1998). Informing teaching, and thereby affecting learning, is one of three main purposes for interim assessment and is the primary use claimed by most school administrators (Clune & White, 2008; Dadey & Diggs, 2019). Common formative assessments also aim to inform teaching and are considered separately from interim assessments because they are often teacher developed (Heredia et al., 2016).

Classroom Formative Assessment

Several authors (e.g., Harlen, 2005; Penuel & Shepard, 2016; Ruiz-Primo & Brookhart, 2018; Shavelson et al., 2008) have suggested typologies for classroom formative assessment. Wiliam and Thompson (2008, p. 63) presented the formative assessment framework depicted in Table 15.1.

The framework shows how five key formative assessment strategies are related to the formative assessment process and its agents. The formative assessment process, sometimes called the formative assessment cycle, can be expressed as a question

Table 15.1 Framework Relating Formative Assessment Strategies to the Formative Assessment Process and Its Agents

	Where the Learner Is Going	Where the Learner Is Right Now	How to Get There
Teacher	Clarifying learning intentions and sharing criteria for success	Engineering effective classroom discussions and tasks that elicit evidence of learning	Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	Activating students as owners of their own learning	

Note. From “Integrating Assessment With Learning: What Will It Take to Make it Work?”, by D. Wiliam and M. Thompson, 2008, in *The Future of Assessment: Shaping Teaching and Learning* (1st ed.), pp. 53–82, C. A. Dwyer (Ed.), Routledge. Copyright 2015 by Taylor & Francis Group. Reprinted with permission.

loop from the point of view of students (Where am I going? Where am I now? How will I get there? [or sometimes How will I close the gap? or Where to next?]) (Assessment Reform Group, 2002; Hattie & Timperley, 2007; Sadler, 1989). This process is a scaffolded, practical presentation of the phases of the regulation of learning (e.g., forethought, performance, and self-reflection; Zimmerman, 2011). The phases overlap; for example, students can use information from any phase to do additional planning (forethought) or to further clarify the learning target in their minds (Where am I going?).

As the framework shows, student involvement is built into formative assessment. Regulation of learning can be taught and can begin as early as preschool. Self-regulation skills are important for all learners, including students with learning disabilities (Butler & Schnellert, 2015). This kind of formative assessment, where students understand what they are meant to learn and regulate their learning in pursuit of these learning intentions, helps students make sense of their own learning (Cizek et al., 2019; James, 2017). Co-regulation of learning (Allal, 2010, 2011; Andrade, 2013; Andrade & Brookhart, 2020; Hadwin et al., 2011) or interactive regulation (Perrenoud, 1998), where regulation of learning is affected by both self and other sources, is invoked when students use external information, for example, feedback from teachers or peers or information from print or electronic sources, which happens in most, if not all, classrooms.

The classroom context affects formative assessment practices. For example, in an evaluative classroom context, where it is not safe for students to be wrong, assessment can feel like quizzing to students because teachers use the information to categorize and judge them (Ames, 1992). In a classroom context more focused on learning than evaluating, where mistakes are seen as opportunities to learn, students are more likely to perceive information, even criticism, as helpful and positive (Brookhart, 2018; Jonsson & Panadero, 2018).

PURPOSE AND USES OF CLASSROOM FORMATIVE ASSESSMENT As Table 15.1 suggests, the central purpose of classroom formative assessment is to inform students' studying and learning and teachers' instruction and is characterized by a series of questions. Student use of information can range from very specific (for example, as students learn to use a capital letter and a period in a sentence, they may use a simple checklist) to more complex (for example, as students learn to organize essays, they may use a rubric to evaluate the presentation of their argument).

The two primary ways teachers use classroom formative assessment are to provide effective feedback about strengths and next steps and to decide on next instructional moves. Therefore, the main goal is not measurement of student achievement of a learning goal; rather, the primary information sought is a description of the status of student understanding (Heritage & Heritage, 2013). What is the student thinking? What is the next thought or experience the student should have? Some

descriptive evidence suggests that teachers who are expert in formative assessment use student work to understand how students are thinking rather than how correct their answers are (Davis, 1997; Hattie, 2009; Hattie & Timperley, 2007; Kroog et al., 2014; Minstrell et al., 2010; Otero, 2006). Information about student thinking is useful for supporting targeted next instructional moves or feedback (for example, “You seem to have more trouble when the fractions need simplifying; let’s work on that”).

DESIGN OF CLASSROOM FORMATIVE ASSESSMENT As Table 15.1 shows, Wiliam and Thompson (2008) described five strategies for formative assessment: (a) clarifying learning targets and criteria for success with students, (b) engineering effective discussions that elicit evidence of student learning, (c) providing feedback that helps improve learning, (d) activating students as resources for one another, and (e) activating students as the owners of their own learning. Table 15.2 gives some examples of tools and tactics for each strategy. Notice that some of these are informal and embedded in instruction and others look like more conventional, and more formal, assessment instruments. What makes these examples of formative assessment is that they are used formatively.

Probably the most noticeable, and the most foundational, change in teaching for many of those who practice formative assessment in the classroom is the explicit sharing of learning goals and the success criteria that will indicate how students are learning. All formative assessment practices depend in some way on that foundation (Brookhart, 2020; Wiliam & Thompson, 2008). For example, feedback from teacher, self, or peers may be based on the criteria. Students’ goal setting and self-regulation are based on their nascent understanding of what they are trying to learn. Questioning and other checking-for-understanding strategies monitor student progress toward the learning goal.

Feedback in Formative Assessment While all assessments provide some sort of feedback, feedback has a special place in formative assessment. Some definitions of formative assessment equate it with feedback. Other definitions of formative assessment consider feedback a major part of formative assessment but not the whole of it (Ruiz-Primo & Brookhart, 2018).

Originally, the study of feedback had its roots in behaviorism (Kluger & DeNisi, 1996), and feedback was seen as a type of reinforcement. Kulhavy (1977) critiqued this idea, pointing out that the stimulus–response environment in any learning program is constantly changing, and there is no good reason to believe that the same sequence of events that the feedback is supposed to reinforce will occur again. Further, he pointed out that feedback that is delayed for a day or more typically leads to increases in student retention, and if feedback were acting as a reinforcer, the effect would decay, not increase.

Table 15.2 Examples of Classroom Formative Assessment

Formative Assessment Strategy	Examples in Practice
Clarify learning targets and criteria for success	• Share and unpack learning target statements with students • Share and discuss examples of good work (exemplars) or examples of a range of quality of work with students • Create checklists, rubrics, or other tools that organize criteria and have students use them to monitor their learning • Co-construct criteria with students
Engineer effective discussions and tasks that elicit evidence of student learning	• Use open-ended and strategic teacher questions • Organize student response patterns so they respond to each other rather than to the teacher • Teach students how to ask effective questions • Check for understanding during learning: ◦ hand signals ◦ whiteboards ◦ quick writes ◦ concept maps ◦ question boxes or boards ◦ entrance or exit tickets ◦ student response systems with multiple-choice questions ◦ quizzes ◦ performance assessment ◦ structured discussions ◦ predict–observe–explain tasks (Shavelson et al., 2008)
Provide feedback that helps improve learning	• Using criteria, comment on strengths and next steps • Written feedback • Oral feedback and dialogue • Video feedback • Give students opportunity to use the feedback
Activate students as resources for one another	• Peer feedback
Activate students as owners of their own learning	• Self-assessment • Reflection at pause points during learning or at the end of an episode of work

Note. Adapted from “Integrating Assessment With Learning: What Will It Take to Make it Work?,” by D. Wiliam and M. Thompson, 2008, in *The Future of Assessment: Shaping Teaching and Learning* (1st ed.), pp. 53–82, C. A. Dwyer (Ed.), Routledge.

In recent years, feedback has come to mean “information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding” (Hattie & Timperley, 2007, p. 81). Interestingly, early studies of the feedback designed and executed using a behaviorist paradigm showed little to no effectiveness on learning (Harris & Rosenthal, 1985; Kluger & DeNisi, 1996), while more recent studies of feedback designed and executed using more constructivist approaches show much larger effects (Graham et al., 2015; Hattie & Timperley, 2007; Kluger & DeNisi, 1996). However, the effects of feedback are quite variable, and—as for formative assessment in general (Klute et al., 2017)—some are negative (Kluger & DeNisi, 1996). Studies with the highest effect sizes (Hattie & Timperley, 2007, p. 84) “involved students receiving information feedback about a task and how to do it more effectively. Lower effect sizes were related to praise, rewards, and punishment.” Finally, it may be as important—or even more important—to study the kinds of response processes feedback engenders as it is to study the effects of feedback on performance (Kluger & DeNisi, 1996; Van der Kleij & Lipnevich, 2021; Winstone et al., 2017).

In other words, some types of feedback are more powerful than others. Outcome feedback, sometimes called knowledge of results or verification (Shute, 2008), is the simplest and most common type of feedback (Butler & Winne, 1995). Outcome feedback tells students whether they were correct or incorrect. Formative assessment can produce this kind of feedback, and it is useful for some purposes, especially for tasks involving recall (Mason & Bruning, 2001). Outcome feedback about correctness coupled with knowledge of the correct response, as, for example, on the back of math fact flash cards or in some computer learning software, can be effective for memory tasks.

In contrast, cognitive feedback, sometimes called elaboration (Shute, 2008), contains information that students can use in their thinking (Butler & Winne, 1995; Hattie & Timperley, 2007; Tunstall & Gipps, 1996). Cognitive feedback helps students interpret the task, interpret their response or response processes, set goals and monitor progress, address particular errors, and give examples or guidance. These are the processes of the formative learning cycle, expressed in the language of self-regulation. This kind of feedback generally has more powerful effects on learning (Hattie & Timperley, 2007; Shute, 2008), although it is important to note that most of the studies in these reviews focused on current performance, not long-term learning (Soderstrom & Bjork, 2015). Teacher pedagogical questioning may be the most effective and immediate elaborated feedback, co-constructed with students in dialog (Heritage & Heritage, 2013). The distinction between outcome and elaborated feedback proves important in computer learning as well (Mason & Bruning, 2001; Van der Kleij et al., 2015). Of course, the possibility of elaborated feedback is related to the cognitive demands of the assessment task; complex tasks typically provide more opportunity for elaborated feedback.

Finally, personal feedback (e.g., “You’re so smart!”) is typically not effective for learning (Hattie & Timperley, 2007). This feedback does not provide information for improvement and may depress motivation.

Koenka et al. (2019) performed meta-analyses of studies investigating the impact of grades (outcome feedback) versus no feedback and grades versus comments on academic motivation and achievement in elementary and secondary school, respectively. Overall, receiving grades positively influenced achievement but negatively influenced learning, compared with no feedback. Overall, receiving comments as feedback resulted in higher achievement and internal motivation compared with receiving grades. Moderator analyses identified more complex relationships. Receiving grades demonstrated no impact on internal motivation for high achievers, as compared with no feedback, but was associated with lower internal motivation for lower achievers. Task-focused comments led to higher achievement and internal motivation, while global-affective comments (e.g., “Excellent! Keep it up!”) resulted in a smaller increase in internal motivation and no difference in performance. In general, these analyses supported the notion that outcome feedback can be helpful; however, for most learning tasks, especially performance tasks that require more than recall, elaborated feedback is more effective (Koenka et al., 2019).

From a measurement perspective, what connects outcome and elaborated feedback is the fact that all types of feedback are—or should be—based on the intended learning for which one seeks information (Brookhart, 2020). The intended learning outcome functions as the construct in assessment to inform teaching and learning (Haertel, 1985). From a learning theory perspective, what connects outcome and elaborated feedback is that both can be informational to students engaged in the self-regulation of learning and to teachers who wish to facilitate student learning in their instruction.

Student Use of Feedback If feedback is to affect learning, students need to make sense of it and use it productively. To do this, students need to appreciate feedback; make judgments about the quality of their own work; manage affect, especially negative emotional responses; and take action that improves their work. The first three of these skills interact, and all three affect students’ abilities to take action to improve their work or learning strategies (Carless & Boud, 2018). One important aspect of managing affect involves student beliefs about whether ability is fixed or malleable; the latter belief is known as a growth mindset. Yeager et al. (2019) found that a brief, online growth mindset intervention reduced the prevalence of fixed mindset beliefs among lower achieving ninth graders, increased grade point average in core classes, and was more effective in schools where peer norms were supportive of a growth mindset—an important finding that highlights the importance of the relationship between individual beliefs and the classroom climate.

Winstone et al. (2017, 2019) named this constellation of skills and dispositions learners' *proactive recipience* for feedback. They identified four recipience processes: self-appraisal, assessment literacy, goal setting and self-regulation, and engagement and motivation. Students' use of feedback depends on these processes as well as the characteristics of the feedback message and its content and the characteristics and behavior of the sender and receiver of the message (Winstone et al., 2017).

Jonsson (2013) found five challenges to university students' productive use of feedback. First, feedback needs to be useful. For example, one-liners and personal comments may not provide information students can use. Second, students prefer specific, detailed, individualized feedback, in part because they perceive it as helpful to them (e.g., Ferguson, 2011; Gamlem & Smith, 2013). Third, authoritative feedback may not be helpful, and much feedback in higher education has an authoritative tone. Fourth, many students do not know how to use feedback productively and may use it passively or indirectly (e.g., making a mental note for next time) or not at all, instead of revising and improving current work. Fifth, many students may not understand the feedback they are given, either because the writing is illegible (which Jonsson, 2013, found to be "a surprisingly common problem," p. 69) or because they do not understand the academic vocabulary, concepts, or criteria in the teacher's feedback. Jonsson and Panadero (2018, p. 546) extracted three conditions for productive student use of feedback from the growing literature on proactive recipience: (a) it helps if students perceive that the feedback is useful, (b) students need strategies for using the feedback, and (c) the feedback should be delivered without a grade. Similarly, writing about the K–12 context, Duschl and Gitomer (1997) argued that frequent, diagnostic assessment activities in classrooms can help students develop the thinking, reasoning, and problem-solving skills necessary for developing science learners.

In summary, feedback is integral to the learning process from both teachers' and students' points of view. Formative use of feedback is enhanced by teacher clarity and elaboration and students' understanding and productive use of feedback.

MEASUREMENT CONSIDERATIONS IN CLASSROOM FORMATIVE ASSESSMENT The validity or trustworthiness of information from formative assessment may be evaluated differently depending on the type of assessment. Informal formative assessment is usually based on some form of classroom questioning (e.g., questions used to prompt classroom discourse, for exit tickets or student reflection exercises, and so on). More formal formative assessment is based on a more formal instrument or tool, for example, an embedded performance assessment or quiz at a hinge point in a sequence of lessons.

A fairly robust literature on the quality of classroom questions predates the arrival of formative assessment in common practice (Cotton, 1988). In general, instruction that included classroom questioning during lessons produced higher achievement than instruction without classroom questioning. On average (Cotton, 1988),

approximately 60% of questions asked required lower cognitive processes to answer (e.g., recall or simple comprehension), 20% required higher cognitive processes, and 20% were procedural questions. Questions requiring lower cognitive processes were more effective at building factual knowledge and memory (Cotton, 1988). However, for older students, increases in questions requiring higher cognitive processes were related to increases in on-task behavior, length of student responses, number of relevant volunteered contributions, number of student-to-student interactions, student use of complete sentences, speculative thinking on the part of students, and relevant student questions.

More recently, researchers have attended to the quality of classroom discourse, a slightly broader notion than classroom questioning. Allowing students to talk about their thinking surfaces that thinking for interpretation by teachers, peers, and the students themselves (Michaels et al., 2008), so that it can be evidence for future teaching and learning moves. Enabling productive classroom discourse involves posing a question that is relevant to the intended learning goal and extending student discourse with prompts that help students respond to each other's thinking, for example, asking a student to restate someone else's reasoning, add their own reasoning, and explain their thinking.

Ruiz-Primo and Furtak (2006, 2007) found a relationship between the quality of middle school science teachers' informal questioning in formative assessment and the amount of student learning, although their study did not allow conclusions about causality. Students learned more in classrooms where teachers deliberately elicited student responses with the express purpose of using the information to move students closer to learning goals and interpreted and acted on the information accordingly. Classroom discourse also socializes students into the classroom learning community. For example, Forman et al. (2017) showed how the use of classroom discourse in a high school science classroom moved a classroom with more didactic question-and-answer discourse toward the kind of scientific argumentation more closely resembling that of scientific communities.

These kinds of projects do not shed light on the quality of teacher-developed formal formative assessment tools like classroom quizzes and classroom performance assessments that are authored or selected by teachers. Earlier literature, mostly from the 1980s and 1990s, evaluated the quality of teachers' classroom testing (Marso & Pigge, 1993), but the authors are not aware of current studies of the quality of the tests teachers use for formal formative assessment outside research projects. Similarly, despite the fact that calls for teachers to use more assessments of authentic performance are not new (Darling-Hammond et al., 1995; Wiggins, 1998), the authors are not aware of current studies of the quality of teachers' performance assessments, which may trade off broad representation of an achievement construct for depth in one area.

IMPLEMENTATION OF CLASSROOM FORMATIVE ASSESSMENT IN PRACTICE Most definitions of formative assessment emphasize the process of implementation, not the validity or trustworthiness of the information (Wylie, 2008). Teacher inferences

from student answers to questions can be key to making on-the-spot decisions to further student learning. Heritage and Heritage (2013, p. 176) described this interactional pedagogical practice as “the epicenter of instruction and assessment.” In a formative teaching environment, assessment by a teacher can be likened to hypothesis testing. A teacher can make a tentative inference about student understanding from a question and then revise that hypothesis as new information comes in, leaving much more room for revising inferences than for assessment that produces a static score report.

Teacher Use of Formative Assessment Practices While there are some isolated bright spots, on balance, research on teachers’ use of formative assessment information to provide formative feedback shows it is infrequent (Johnson et al., 2019; Schneider & Andrade, 2013; Yan et al., 2021). Some evidence suggests that interactional instructional decisions and follow-through that truly target a student’s current level of development and move the student toward deeper understanding for a particular concept are difficult to do and that high-quality practices of this type occur rarely and require rich professional development (Furtak et al., 2008; Heritage & Bailey, 2006; Randel et al., 2016; Wylie & Lyon, 2015).

When teachers shift from a view of teaching that emphasizes what they will present to a view of teaching that emphasizes what students will be trying to learn and what they will do to learn it, the classroom learning climate necessary to support the full benefits of formative assessment can be realized. This shift is a major change, from “instructivist” (Box et al., 2015, p. 972) or “social efficiency” (Shepard, 2000, p. 4) teaching focused on delivering content to learners to teaching that supports students as they construct their own meaning. Many teachers find this shift difficult. Research results about the effectiveness of preservice and in-service teacher education in formative assessment are mixed (Brookhart, 2017). In general, at both preservice and in-service levels, where attention was paid to changing teachers’ beliefs about learning and those beliefs did in fact change, there was evidence that teachers improved their formative assessment practices. These successful examples function as proofs of concept that effective formative assessment processes can be taught, but such developments are not common enough to be the rule. Other reviews of literature about teaching formative assessment practices and processes have come to similar conclusions about the necessity for, and difficulty of, teachers moving from a teacher-centered to a student-centered understanding of teaching and learning for formative assessment to be effective (DeLuca, 2012; Otero, 2006).

More than any other type of assessment, effective classroom formative assessment is advanced when teachers know both the principles of effective assessment (for example, how to construct questions and tasks that elicit evidence of learning) and how to observe or score the results. Similarly, effective classroom formative assessment is advanced when students know how to use assessment information to inform their learning. Effective classroom formative assessment also advances with the understanding and support

of administrators, especially building principals (Michigan Assessment Consortium, 2017; Schneider & Andrade, 2013).

Even when questions and tasks do elicit evidence of student thinking, it appears that it is difficult for teachers to use the information appropriately (Heritage et al., 2009; Ruiz-Primo & Furtak, 2006, 2007; Ruiz-Primo & Li, 2013; Schneider & Gowan, 2013). Despite the difficulty, professional development can increase teachers' abilities to design tasks and questions that elicit student thinking, interpret the resulting information, and use it to feed students' learning forward (Furtak et al., 2016).

Ruiz-Primo et al. (2015) explored the type of teacher actions resulting from judgments based on on-the-fly, informal formative assessment. They analyzed videos from 20 science and mathematics teachers and found the same patterns in each discipline. Meaningful interactions resulting from on-the-fly formative assessment were observed in 67% of the mathematics instructional tasks observed and 71% of the science tasks, sometimes also coupled with interactions to keep students on task. However, both mathematics and science teachers responded to formative assessment information with verbal feedback more than with instructional moves, and most of these responses occurred in the context of whole-class instruction as opposed to the context of individual work. These are only a few of the response patterns teachers could have at their disposal.

Student Use of Formative Assessment Practices Students' use of formative assessment varies. For example, students need to be taught to be effective peer assessors (Topping, 2013; van Zundert et al., 2010). Classroom and school norms affect students' formative beliefs (Yeager et al., 2019) and therefore their use of formative information. Boekaerts et al. (2006) emphasized that students pursue multiple goals in the classroom simultaneously, including wanting to be entertained, to belong, and to feel safe and valued, in addition to their academic learning goals, and that the instructional climate affects each of these. Similarly, Leighton et al. (2013; Leighton, 2019) proposed that the instructional climate, learners' mental models of emotion and cognition, and students' performance—their use of feedback—are related in predictable ways.

IMPACT OF CLASSROOM FORMATIVE ASSESSMENT ON LEARNING As students develop greater facility in regulating their learning, they become more powerful learners. As Black et al. (2006, p. 126) summarized, "Pupils' learning is more productive if it is reflective, intentional, and collaborative, practices which may not come naturally but which can be taught and can lead to pupils taking responsibility for their learning." Students who understand what they are trying to learn can be more intentional about their learning than those who do not, who simply comply with teacher directions. Looking at examples of work is one way that students come to develop an understanding of what high-quality work looks like and, by inference, the understandings and skills that underlie high-quality work. As Sadler (1989, p. 121) wrote, "In other words, students

have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it.”

Increased achievement and increased student self-assessment capabilities result when students understand the criteria for good work (Andrade & Valtcheva, 2009). This has been found in many subject areas and grade levels, for example, primary projects (Higgins et al., 1994), elementary and middle school writing (Andrade et al., 2008, 2010; Coe et al., 2011), middle school mathematics (Ross et al., 2002); middle school special education (E. Lee & Lee, 2009); secondary school social studies (Panadero et al., 2012; Ross & Starling, 2008), and college biology (Hafner & Hafner, 2003).

Reviews of research on rubrics, which are one way to codify and communicate criteria for students, have suggested that one of their main functions is to clarify expectations and quality criteria (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). It is important that the criteria promote understanding and not a mechanical criteria compliance for a short-term performance boost (Balloo et al., 2018; Gitomer & Duschl, 1996; Sadler, 2014).

Estimating the size of the effect of formative assessment on student learning has proved more difficult than it sounds because formative assessment, as the framework in Table 15.1 shows, encompasses so many different strategies. In addition, fidelity of treatment can be an issue. Meta-analyses of the effects of formative assessment on achievement have shown generally positive effects (Kingston & Nash, 2011; Klute et al., 2017; H. Lee et al., 2020; Wiliam et al., 2004), with some of the effects being negative. Studies of the effects of formative assessment practices in specific disciplines have also found generally positive, but variable, effect sizes (Andrade et al., 2019; Burkhardt & Schoenfeld, 2019; Decristan et al., 2015; Graham et al., 2015; Herman et al., 2015). The variability may reflect the fact that formative assessment studies are not necessarily all investigating the same thing. When studies are screened to focus on a single important aspect of formative assessment, effect sizes can be higher. For example, H. Lee et al. (2020) found that formative assessment interventions that featured student-initiated self-assessment had an effect size of 0.61. Hattie and Timperley (2007) found an average effect across meta-analyses of feedback studies of 0.79.

EVALUATION OF CLASSROOM FORMATIVE ASSESSMENT INFORMING TEACHING AND LEARNING Classroom formative assessment is the type of assessment most clearly associated with informing teaching and learning. While elements of it such as feedback have a long history, classroom formative assessment has made a relatively recent arrival into assessment research. Theory is in development (e.g., Table 15.1) and research agendas have begun.

Classroom formative assessment, especially the more informal types, stretches conventional measurement thinking in terms of the nature and quality of the assessment process and tools. This stretching is necessary to bring the benefits of sound

assessment squarely into the learning community, typically a classroom, where learning takes place and where students and teachers can gather and use information to inform teaching and learning as it happens. Each classroom is a unique community of learners based on a give-and-take between a particular teacher or coteachers and their students. Classroom information requirements differ from the information requirements for standardized assessment—where context should not matter—in important ways. Formative assessment foregrounds the role of the student and often invokes contemporary learning theories that view the student as an active participant in assessment and learning. This changes the role of students from test takers to learners. Formative assessment is heavily dependent on the quality of the process, not just the quality of the tools, instruments, or questions, for its effectiveness. Formative assessment works best when both students and teachers know how to use it effectively. While research suggests formative assessment can have more impact on teaching and learning than other types of assessment, these challenges and others must be met, even if gradually, for classroom formative assessment to reach its full potential.

Interim and Benchmark Assessment

Many educators and districts use the terms interim and benchmark assessments as synonyms, and the terms have been used interchangeably in the research literature as well. Sometimes a distinction is made: Interim assessments can be parallel test forms for an external accountability test, covering an entire year's worth of content and administered two or three times during the school year to track student learning and achievement growth; and benchmark assessments can be nonparallel test forms covering a portion of the year's content (e.g., the first report period) and administered at a specified point in the school year and curriculum (Ferrara et al., 2019). This chapter uses the term interim assessments for both.

Many interim assessments are sold by test vendors. Interim assessments can also be sold as item banks or as item banking and testing software that can be populated with locally written items or items from other sources. Burch (2010, p. 147) defined interim assessment technology as “the software that is sold to schools and districts in order to gauge students' progress towards high-stakes summative tests and to comply with the test reporting requirement of the No Child Left Behind Act (NCLB) of 2001.” She argued that the upsurge in use of interim assessment technologies represents an intensification of educational privatization brought on by NCLB (Burch, 2006, 2010) and that this is at least partly because of pressure on schools to be “more efficient, more compliant, and more equitable” since NCLB (Burch, 2010, p. 147).

PURPOSE AND USES OF INTERIM ASSESSMENT Interim assessments are used to serve one or more of three general purposes: instructional, predictive, and evaluative (Perie et al., 2009). For example, an instructional purpose might be to identify particular

content or standards for remediation or reteaching. A predictive purpose might be to determine each student's likelihood of attaining a proficient score on the end-of-year state accountability test. An evaluative purpose might be to provide information about the effectiveness of an instructional program or some curriculum materials. In practice, most districts use interim assessments for several different purposes simultaneously, most commonly for instructional and evaluative purposes (Dadey & Diggs, 2019; Davidson & Frohbieter, 2011). Even when districts use interim assessment for predictive purposes, the intent of the prediction is to inform instruction before the state accountability test, not simply to make the prediction.

DESIGN OF INTERIM ASSESSMENT As the above definitions imply, most interim assessments were designed primarily to give predictive information. They often use a small sample of items across many standards and at the standard level are most reliable for decisions about groups. However, the logic used by district administrators when they purchase interims is that an investment in gathering interim information will support alignment of teaching and learning in the district, provide practice for state accountability tests, and provide information for teachers on the needs of both current and future students (Clune & White, 2008)—again illustrating that trend to want to add a student-level diagnostic purpose to all assessments. The fact that one assessment would not serve these purposes equally well is not typically considered (Perie et al., 2009); however, it is an important consideration in evaluating research on the effectiveness of interim assessments (Immekus & Atitya, 2016).

Interim assessments may be computer-administered or paper-and-pencil tests and usually result in one or more of the following kinds of scores: scale scores, percentiles, and proficiency-level ratings. The results are meant to be used by teachers and administrators to assist individual students (for grouping, remediation, and so on) or to assist with group planning (to identify strengths and weaknesses in current instruction and plan for future instruction). The intention is that educators will set goals, measure and evaluate progress, and use data to make changes. This process is focused mainly on educator learning and is relatively silent on student learning. The intended connection between interim assessments and student learning comes from the claim that measures are linked to learning standards. The most common responses to interim data are reteaching and review (Penuel & Shepard, 2016).

MEASUREMENT CONSIDERATIONS IN INTERIM ASSESSMENT Some commercially produced interim assessments have technical manuals that present standard kinds of information about reliability, validity, and scaling (e.g., Northwest Evaluation Association, 2019). Some, of course, do not. Many of the more established commercial products are reviewed in the Buros Mental Measurements Yearbook series (<https://buros.org/mental-measurements-yearbook>). Validation differs depending on the intended use of the interim: whether instructional, evaluative, or predictive (see also Lane & Marion, this volume).

Arguably the predictive use requires the simplest and most readily available validity evidence, to answer the question of how well performance on the interim predicts state accountability test scores. R. S. Brown and Coughlin (2007) investigated the availability and quality of predictive validity information for a selection of interim assessments used in the Mid-Atlantic region. They considered four assessments (the Northwest Evaluation Association's Measures of Academic Progress, MAP; Renaissance Learning's STAR; Study Island; and CTB's Terra Nova). While all four provided concurrent or predictive validity information of some sort, only one, Terra Nova, presented a truly predictive study.

Pereira and Tienken (2012) undertook a study in four middle schools to support the validity of using a computer-based interim assessment tool as recommended by the New Jersey Department of Education. The tool had pre- and posttests in mathematics and language arts. The product was marketed to predict achievement on the state accountability test. Findings indicated that the pretest and posttest had almost identical predictive power, prompting the authors to question the need for a posttest (Pereira & Tienken, 2012, p. 11). Further, Babo et al. (2014) reported that the odds ratios for passing the state test for pretest and posttest scores were close to 1, indicating little predictive power. This finding is consistent with broader finding of the large contribution of general mental ability to assessments of educational achievement (Deary et al., 2007).

IMPLEMENTATION OF INTERIM ASSESSMENT IN PRACTICE Dadey and Diggs (2019) identified and summarized 20 studies of interim assessment use. They found the most frequent reported uses were instructional, usually in the form of broad claims that interim results would be used to modify or improve instruction. All 20 of the studies they reviewed listed instructional purposes, for example, general instructional planning, grouping students, and providing remedial support. Sixteen studies also mentioned evaluative purposes, citing evaluation of teachers, programs, and curriculum and supporting resource allocation at the staff, school, and program levels. Eight studies cited predictive use of interim results and always the same one: predicting student performance on the state accountability test.

The process of gathering interim assessment information and using it to make decisions is usually framed as some sort of data-driven decision-making process (Penuel & Shepard, 2016), often more about management than about learning. Acknowledging this shortcoming yet wishing to harness the formative power that can be gained by gathering relevant data for specific learning needs, Wiliam (2018a, p. 47) called for "decision-driven data collection."

Marsh (2012) described a data use theory of action with five steps: (a) access and collect data; (b) turn data into information by organizing, filtering, and analyzing; (c) turn information into knowledge by combining it with teacher understanding and expertise; (d) apply the knowledge to arrive at a response and action, and (e) assess the effectiveness of the outcomes of that action. This is a very linear expression of what is inherently an interpretive process (Blanc et al., 2010; Farrell & Marsh, 2016). Accordingly, the

research studies in their Philadelphia project (Blanc et al., 2010; Christman et al., 2009) reframed the teacher data use process as sense making. They described three types of sense making in which teachers engaged with interim data. They described strategic sense making as the most common; it focused on short-term adequate yearly progress targets—things like identifying students who could, with just a little help, become proficient. Affective sense making, also common, focused on teachers' professional responsibilities and things they could do to motivate students and improve learning. Reflective sense making was least common; it focused on questioning and evaluating one's instructional practices.

The Urban Data project of the Council of the Great City Schools (Faria et al., 2012; Heppen et al., 2010, 2011) situated teacher data use within a broader theory of action with five key dimensions: (a) the district's assessment context, (b) supports for data use, (c) working with data, (d) instructional responses, and (e) improved student achievement. This theory of action considers contexts and supports as foundational to data use and achievement outcomes. Heppen and colleagues (2010) found that the assessment context in schools predicted the other key dimensions and that teacher attention to data was a significant, positive predictor of instructional responses. Bulkley et al. (2010) concluded that interims could serve instructional purposes, but only if strong district supports were in place.

The use of interim data is limited by time constraints, especially lack of time for remediation and the presence of pacing guides for instruction (Abrams et al., 2016), and also by the data culture or context in schools (Goertz et al., 2009; Heppen et al., 2010), as well as teachers' own facility with and attention to data (Faria et al., 2012; Heppen et al., 2010). One study using the web logs behind a district's information dashboard (Tyler & McNamara, 2011) found that, on average, teachers spent 2.3 minutes per week looking at the class roster page of their interim results and 30 seconds per week looking at individual student pages—or less than 3 minutes per week on the two results pages that contained information relevant to instructional adjustments. Another study of the use of interim assessments in an urban charter high school found that the tests, based as they were on grade-level standards, did not provide much useful information on students who were often more than 2 years behind in grade-level achievement. In addition, there were too many separate and varied standards on the tests to help teachers make decisions about what to do, since the students needed assistance in most or all of them. For these reasons, that testing program was abandoned after just over 3 years (Bancroft, 2010). Another study, using randomized controlled trials, found little evidence of changes in teacher practices as a result of having access to an interim assessment tool (Chojnacki et al., 2013).

Several studies describe the instructional decisions teachers report making on the basis of interim assessment results, and these studies also note great variability by teacher, including some teachers who do not use the information. Teachers who do use interim results tend to use it to identify topics to teach and students with needs in those topic areas (Abrams & McMillan, 2013; Abrams et al., 2015; Goertz et al., 2009; Oláh et al., 2010; Shepard, 2010), although these decisions are influenced by teachers'

perceptions of the difficulty of particular content and students' past performance (Goertz et al., 2009) and their perceptions of the assessment's design and item quality (Abrams et al., 2015; Farrell & Marsh, 2016). Most teachers have not been found to use interim results to change the way they teach content or to look for underlying reasons students may have developed misunderstandings (Christman et al., 2009; Goertz et al., 2009; Oláh et al., 2010). For example, a study of teachers' use of mathematics interim assessments found teachers mostly used results to group students or reteach procedural knowledge, rather than making sense of students' conceptual understanding (Oláh et al., 2010). Teachers report using interim data to identify students who could potentially move from *Basic* to *Proficient* or *Below Basic* to *Basic* by the time of the state test, to identify skills and concepts to reteach, to identify students with similar areas of weakness or strength, to change some classroom routines (Blanc et al., 2010), and, in mathematics, to locate and diagnose errors (Goertz et al., 2009).

Several studies discussed teachers' use of interim data as part of a larger assessment system that included classroom formative assessment, teacher-developed classroom summative assessments, teacher-developed common formative assessments, interim assessments, and state accountability assessments. Wilkerson et al. (2021) surveyed teachers in Nebraska and found that some teachers did not use data at any level (state summative, interim, or classroom formative), but those who did used data from interim assessments once or twice a month to inform instruction. Farrell and Marsh (2016) found teachers had the most mixed opinions of interim assessment data, which they often found less useful and less trustworthy than other kinds of assessment data. In the span of a district assessment system, data from state accountability tests sometimes influences grouping and instructional decisions at the beginning of the year (Abrams et al., 2016). Interim assessment data, as noted above, is most often used to decide *what* to teach to whom, often looking backward at what content needed reteaching (Abrams et al., 2016; Farrell & Marsh, 2016; Riggan & Oláh, 2011). Classroom formative assessment, in contrast, is used to help get students to explain their thinking (Abrams et al., 2016; Riggan & Oláh, 2011) and support decisions about *how* to teach or reteach going forward. Sometimes teachers looked to classroom formative assessment information to help explain or deepen their understanding of interim assessment results to make instructional decisions (Martone et al., 2018).

To use interim assessments well, teachers and administrators need to be able to read, interpret, and use a wide range of both norm- and criterion-referenced scores, as well as understand the concepts of standard error of measurement and standard error of the mean or other aggregated scores, the limited reliability of change scores, and the like. They also need the skills to communicate assessment results to parents and students in an understandable fashion. Many test vendors provide teacher resource material and professional development as part of their interim programs, but as reported previously, the evidence suggests that the inclusion of these resources does not necessarily lead to deeply conceptual differentiated instruction or to improved student achievement.

IMPACT OF INTERIM ASSESSMENT ON LEARNING A small number of studies have evaluated the effects of using interim assessments on student achievement. Faria et al. (2012) found that teacher data use and student achievement were statistically significantly but weakly correlated in middle school mathematics and elementary reading but not in elementary mathematics or middle school reading. Principal data use was statistically significantly correlated with student achievement only in middle school mathematics.

Three randomized studies were found, each investigating different interim assessments. One looked at the effects of using interim assessments (mCLASS in K–2 and Acuity in 3–8) in the state of Indiana in the 2009–2010 school year on the state ISTEP+ test (Konstantopoulos et al., 2013, Konstantopoulos, Li, Miller, & van der Ploeg, 2016; Konstantopoulos, Miler, van der Ploeg, & Li, 2016). Effects in mathematics were statistically significant in Grades 3–8 (not in K–2); effects were larger in the lower (10th and 25th) quantiles. Effects in reading were smaller and not always significant. The exact effect size depended on the grades included in the analysis and whether the analysis examined intent to treat or treatment on the treated (Konstantopoulos et al., 2013). A second study looked at the effects of a program implemented by the Johns Hopkins Center for Data-Driven Reform in Education, including the use of their 4Sight benchmark tests and associated professional development (Carlson et al., 2011), in 500 schools in seven different states, again using the state tests as the dependent variable. They also found significant weak effects in mathematics but not in reading. A third study (Cordray et al., 2012) looked at the effects of using the Northwest Evaluation Association’s MAP program in 32 elementary schools in Illinois in Grades 4 and 5 reading. MAP teachers were not more likely than control teachers to use differentiated instructional practices in their classrooms, and there was no significant effect on reading achievement in either Grade 4 or Grade 5. Another study in Massachusetts, using a matched comparison design, compared the 2001 to 2006 Massachusetts Comprehensive Assessment System (MCAS) scores for high-poverty middle schools that were in a pilot program using benchmark assessment with the scores of schools that were not. The intervention occurred in 2006, when the pilot schools used interim assessment technology software to develop quarterly benchmark tests. There were no differences between treatment and comparison schools, either in 2006 (Henderson et al., 2007) or 1 year later (Henderson et al., 2008).

EVALUATION OF INTERIM ASSESSMENT INFORMING TEACHING AND LEARNING The popularity of interim assessments grew from a perceived need for more frequent information for educational accountability assessment. However, the popularity of interim assessments and the pressure to use them for diagnostic decisions they are not designed to support presents interesting evidence of the movement toward insisting that assessment inform teaching and learning—the trend toward the formative use of assessment information.

On balance, then, two arguments are supported. First, teachers do not make much effective use of data from interim assessments. It does seem that educators’ use of

interim data can be described by one or more of the posited theories of action for data-driven decision-making. However, these processes alone, at least as used to date, offer only a very partial explanation of what happens as educators use data. Many teachers do not engage in these processes, or they do so with inadequate assessment literacy or with instructional adjustments that are topical or procedural rather than adjustments intended to support students' conceptual change.

Second, there is little evidence that using interim assessments improves student achievement. There are only a small number of studies of the effects of interim testing programs on student achievement. These studies yield only small and intermittent effects or no effects. Studies show teachers use item analysis as a main strategy for decisions about reviewing topics and procedures rather than seeking to understand student thinking, and items are only samples from an achievement domain and not the domain—the achievement standards or learning goal—itsself. Therefore, even if teachers did make better use of interim assessment data, it is not likely that their effectiveness would be dramatically improved.

Common Formative Assessments and Other District-Developed Assessments

Some districts design what they call *common formative assessments* to be administered across classes in a specific subject area and grade level at pause points in instruction, for example, after an instructional unit or series of related units. The intention is typically to inform teacher instructional planning.

PURPOSE AND USES OF COMMON FORMATIVE ASSESSMENTS Ainsworth and Viegut (2006, pp. 2–3) defined common formative assessments as “assessments collaboratively designed by a grade-level or department team that are administered to students by each participating teacher periodically throughout the year.” Teachers use information from common formative assessments for instructional planning for current students and for future planning for when they teach the same subject matter again.

DESIGN OF COMMON FORMATIVE ASSESSMENTS Common formative assessments are usually, although not always, tests and are sometimes performance assessments created or curated by teachers who teach common curricular material in a building or across a district. All of the literature, both professional and scholarly, consulted by the authors assumed quantitative scoring of some kind: right/wrong for selected-response tests and the use of a rubric, typically arranged as a proficiency scale on the standard, for constructed-response test items and performance tasks.

For Heredia et al. (2016), collaborative teacher design is part of the common formative assessment process. Ainsworth and Viegut (2006) added collaborative scoring. Ideally, the use of a common tool for formative assessment will decrease the variations in teacher formative assessment practices and potentially therefore maximize student learning (Heredia et al., 2016, p. 698). Some test vendors produce

assessments they call common formative assessments and market them to school districts, which circumvents the teacher design work and its benefits in terms of developing consensus on both formative assessment (Heredia et al., 2016) and supporting student learning (Furtak, Circi, & Heredia, 2018) and makes these products similar to interim assessments.

Common formative assessments are distinguished from other school-based tests, for example, departmental tests in secondary education, in that their use is intended to inform teacher instructional planning and not student grading. At their best, common formative assessments can inform both instructional planning and teacher professional development (Furtak & Heredia, 2014). Common formative assessments are a cornerstone of the professional learning communities movement (DuFour et al., 2010; Myers, 1996) that has gained momentum in school districts around the United States.

MEASUREMENT CONSIDERATIONS IN COMMON FORMATIVE ASSESSMENTS The professional literature recommends that teachers, typically in grade-level or subject area teams, study collaboratively standards, curriculum documents, and other instantiations of what students should know (e.g., textbooks, examples of student work, or previous student performance data) to establish what they will design their common formative assessments to measure (Ainsworth & Viegut, 2006; DuFour et al., 2010). To date, the authors of this chapter have found no studies of the quality of teacher team–developed common formative assessments. However, the heavy emphasis on teacher teams identifying standards or learning progressions and aligning common formative assessment items and tasks to them, in regard to both content and level of thinking, suggests that the content-related evidence for validity—and a match between the tested and taught curriculum—may be a strength of common formative assessments.

To make sure information from locally created common formative assessments is of high quality and can support instructional decisions, the professional literature simply advises district teams to look at recommendations from assessment experts; released items from other testing programs, including on websites such as that for the National Assessment of Educational Progress (NAEP; <https://www.nationsreportcard.gov/nqt/>); and examples of performance assessments and rubrics from local or other sources (DuFour et al., 2010). District teams are advised to refine assessments after use according to established standards for evaluating assessments (Ainsworth & Viegut, 2006). Given the documented variability of teachers' assessment practices in other areas (see, for example, the section on "Grading"), it is reasonable to expect that this advice will be followed more in some places than in others and that there will be variability in the quality of common formative assessments.

IMPLEMENTATION OF COMMON FORMATIVE ASSESSMENTS The process of implementing common formative assessments may be quite loose or it may be more structured, depending on how teams in any given school operate. It is typically part of a school

improvement process. For example, the Team Learning Process (DuFour et al., 2010, pp. 26–33) calls for a process in six steps: (a) identify essential outcomes all students must learn; (b) create common pacing guides and curriculum maps all teachers will follow; (c) develop multiple common formative assessments; (d) establish a target score for proficiency in each skill on each assessment; (e) administer common formative assessments and analyze results; and (f) celebrate strengths and identify and implement improvement strategies. This is a standards-based approach.

The Formative Assessment Design Cycle (Heredia et al., 2016; Furtak & Heredia, 2014) uses a learning-progressions approach, in five steps: (a) reflect, (b) explore student ideas, (c) design tools, (d) practice using tools, and (e) enact tools. Then the cycle repeats, beginning with reflection; typically, the tool (the common formative assessment) is also revised. Reflection is based on current teaching practices as well as student work. Exploring student work can include looking at aggregated data as well as examples of individual student work, with a focus on student thinking.

When done well, the work involved in aligning content standards with the assessments can serve as professional development for the teachers (Frey & Fisher, 2013). Gallimore et al. (2009) conducted a 5-year quasi-experimental study of teacher professional development using an inquiry protocol and found that teachers' attribution of improved student performance changed from implicating external causes to attributing learning to their own teaching. Part of the protocol involved looking at indicators of student learning, designed by the teachers together.

Assessment literacy requirements for teachers creating tests or performance assessments aligned with standards, to be administered across classes, are high. Teacher assessment literacy rises as part of the process (Furtak et al., 2014, 2016; Heredia et al., 2016), and assessment validity rises because the work is informed by collaboration with teachers and access to their pedagogical content knowledge (Ball et al., 2008). Even with support, however, teacher growth in understanding of learning progressions and the assessment strategies and tools that will elicit evidence about them takes time and involves struggle (Furtak, 2012; Furtak & Heredia, 2014).

IMPACT OF COMMON FORMATIVE ASSESSMENTS ON LEARNING Little published evidence of the impact of common formative assessments on learning was found. Gallimore et al. (2009) reported evidence that when teachers used inquiry and followed a learning problem long enough to understand it, assess it, and address it, student achievement improved. This professional development included the use of some assessments that could be described as similar to common formative assessments. Frey and Fisher (2013) found that the use of common formative assessments helped increase student achievement in writing.

EVALUATION OF COMMON FORMATIVE ASSESSMENTS FOR INFORMING TEACHING AND LEARNING Common formative assessments involve teachers in a collaborative process to check on student progress on learning standards in their curriculum. In principle this

is a sound idea, and there is some evidence that when done well, teachers and students benefit. However, there is also evidence that assessment design is difficult for teachers to do well, and because of recommendations in the teacher professional literature it is likely that many common formative assessments of unknown quality exist. One key to effective use of this type of assessment is increased literacy on the part of teachers regarding the construction and interpretation of assessments. Increased literacy may also help teachers select, rather than design, appropriate assessments.

Assessment Focused on the Summative

When assessment informs teaching and learning, it is being used for a formative purpose. Therefore, it may seem that assessments designed for summative purposes should have no place in this chapter. However, there is evidence that both effective teachers (Wiliam et al., 2004) and successful students (Brookhart, 2001) use information available from summative assessment to inform future planning. Moreover, as assessment trends toward the formative, it is more and more common that educators are expected to use all available data sources, including summative, to inform teaching and learning (Farrell & Marsh, 2016; U.S. Department of Education [U.S. DOE], 2009; Wilkerson et al., 2021).

In effect, because learning is ongoing, most K–12 summative assessments fulfill their immediate summative purpose, but in the larger picture serve more as periods at the end of sentences rather than as the end of a story. This chapter will consider the use of two common forms of summative assessment, grading and state accountability testing, for formative purposes.

Grading

Grading refers to the symbols assigned to individual pieces of student work or to composite measures of student performance on student report cards. A long history of research has led those who study measurement to question the reliability and validity of grades and, sometimes, therefore, to dismiss their importance. However, grades loom large in the educational experience of all students and are the activity on which teachers spend most of their assessment time, at least at present. They are perhaps the most universally used assessment to inform teaching and learning (Brookhart et al., 2016).

PURPOSE AND USES OF GRADING Grading has a wide array of purposes and uses. Grades are used to inform teachers, students, and parents/caregivers about teaching and learning at many different levels, for example, to support next steps in instruction or studying for a small amount of instructional content or to make broader decisions about next year's teaching or course taking. Grades are used to rank and sort students for various placement and selection decisions, for example, as part of decisions about placing students with special needs in classes or programs, qualifying students to participate in extracurricular activities, and college admissions. Grades can be used to evaluate teachers, schools, and educational programs. Informally, students and teachers

use grades to do all sorts of things they were never meant to do, for example, support classroom behavior management or students' sense of self-worth (Covington, 1992; Thomas & Oldfather, 1997).

A recent grading reform in the United States, called standards-based grading, reports student achievement relative to content standards established for a particular grade level. Reformers claim this gives students, parents, and educators more useful feedback (Peters et al., 2017). In other words, in standards-based grading, grades are assumed or intended to inform teaching and learning and that purpose is assumed to be primary. Standards-based grading, therefore, makes explicit that grades are to be used formatively as well as summatively, something that traditional grading practices typically do not do. Standards-based grading principles call for reporting academic achievement separately from behaviors, effort, and attendance; referencing both assessment and grading to standards; using report card grades to report current achievement status by prioritizing the most recent evidence of learning; allowing students to edit and resubmit work; using proficiency-based rubrics and decision rules for aggregating them that take their ordinal nature into account; and using quality formative assessments as well as summative assessments (Guskey & Bailey, 2001; Peters et al., 2017).

DESIGN OF GRADING Grades come in many forms. For traditional report cards in the United States, the most common is an ordinal scale of letters (often ABCDF). Many high schools report grades as percentages (0–100), and many schools at all levels that use the ABCDF scale for report cards grade individual pieces of student work on the percentage scale. Many other countries (e.g., Botswana, Canada, Sweden) use some version of a categorical scale, typically letters. An ordinal scale of proficiency levels (e.g., *Advanced*, *Proficient*, *Developing*, *Emerging*, or 4, 3, 2, 1) is often used with standards-based grading, a practice that evaluates student achievement against individual standards rather than subject areas overall. The process of assigning grades is at present mostly a teacher function, although students can be involved, and there is evidence students are fairly accurate in their appraisals of their own work (Sanchez et al., 2017).

MEASUREMENT CONSIDERATIONS IN GRADING Grades have the distinct disadvantage that the general public thinks they understand their meaning (“I’ll give that restaurant an A+”; “C is average”), when often such meanings no longer apply, if they ever did. Different stakeholders may view grade categories differently. For example, Waltman and Frisbie (1994) found that on average—and with much variation—parents and fourth graders thought the average mathematics grade was a C+ when in fact it was a B.

Validity In the early 1900s, in what would now be called validation research, studies sought to describe the relationship of grades to intelligence, assuming that intelligence was the construct grades were supposed to measure, just as the purpose of schooling was to sort students so that the bright students received more education and the duller

students left school for more menial jobs (e.g., Banker, 1927). As the philosophy of teaching shifted from sorting students to a more achievement-based approach, studies of the validity of grades shifted to comparing grades with tested achievement, as measured by a standardized test (e.g., Moore, 1939), expecting strong relationships between these two measures of achievement. However, this is not and has never been the case (Bowers, 2011; Brookhart, 2015; Brookhart et al., 2016). Over half a century ago, Miner (1967) demonstrated that graded achievement and tested achievement are not the same construct.

There is ample evidence that teachers use nonachievement factors in grades (Brookhart, 2013), against the recommendations of measurement professionals (Stiggins et al., 1989), which may explain part of the difference. Bowers (2009) termed the variance in grades not accounted for by standardized tests of achievement a “success at school factor” (p. 623). His study replicated the existence of the nonacademic factor in U.S. students’ grades that Lekholm and Cliffordson (2008) had found in Sweden and named it, arguing that this measure of success at school should be useful for educational decision-making beyond the information available from standardized test scores.

That high school grades are a slightly better predictor of first- and second-year college grades than ACT scores (Westrick et al., 2015) also supports the notion that high school grades include some sort of success-at-school factor. Galla et al. (2019) showed that the incremental predictive validity of high school grades (beyond test scores) for on-time graduation from college was explained by student self-regulation, while the incremental predictive validity of test scores (beyond grades) was explained by cognitive ability. This study is important because it adds a theoretical framework, self-regulation, to the practical idea of a success-at-school factor. (See also Camara et al., this volume, for a review of research related to admissions and predictive validity.)

There is some evidence that grades can yield higher quality information if certain effective grading practices are followed (Pollio & Hochbein, 2015; Welsh et al., 2013; Willingham et al., 2002). There is also ample evidence that high-quality grading practices are not always followed (Brookhart et al., 2016). Cumulative grades, but not individual grades, are valid for purposes of predicting and sorting (Anderson, 2018). Evidence is not strong for the validity of either cumulative or individual grades to provide information about student achievement of specific learning outcomes.

Reliability In the early 1900s, the accuracy of the 100-point scale was questioned (e.g., Starch, 1913), and categorical grading was adopted to increase reliability—or at least as a trade-off between perceived precision and reliability. Most grades currently, in the United States and around the world, use a categorical scale of some length. The categories may be based on cutoffs on a percentage or other point scale or on performance-level descriptions for each of the categories.

Recently, attention has turned to the reliability of grade point averages. Beatty and colleagues (2015) found an overall estimate of reliability of .86 for first-year college grade point average and .93 for overall grade point average. Westrick (2017) reported

the range of reliability values for fourth-year cumulative grade point averages at each of 26 four-year institutions was .89 to .92. Thus, while grades on single assessments may be unreliable, cumulative grades are reasonably reliable (Anderson, 2018). Again, though, the reliability in question is reliability for relative ordering. Although this is an important quality, it pertains more to some of grading's many purposes than to others. Reliability of proficiency classification, for example, was not studied.

For some grading purposes, most notably the sorting and ranking ones, the added precision that comes with aggregating, especially up to the level of a grade point average, adds reliability. For other grading purposes, most notably the feedback-related purposes of conveying information about student achievement of specific subject area content or standards, the added precision may muddy interpretation. In the case of a borderline grade, for example, there may not be a meaningful difference in achievement between a very high B and a very low A, but the categories proclaim such a difference and students and teachers may act on the basis of that difference.

When categorical (ordinal) grades are used, nonparametric statistical techniques are best suited to aggregating component grades on students' tests, projects, and other summative assessments into composite grades for reporting. In the early 21st century, that aggregation is most often accomplished with grading software, much of which was developed by programmers who have not studied the nature of grades and treat them parametrically (e.g., using means, standard deviations), with calculations often carried to several decimals and imputing a precision that was not there to begin with. For better or worse, grading software removes some of the decision-making about grading from teachers' hands.

IMPLEMENTATION OF GRADING IN PRACTICE Research on teacher grading practices, for example, what assessments teachers count in a report card grade, has been much studied since the early 1900s and reviewed periodically (Brookhart, 1994, 2013; Brookhart et al., 2016; A. D. Crooks, 1933; Kirschenbaum et al., 1971; Smith & Dobbin, 1960). All these reviews concluded that the most effective grading system would be to report student achievement on established standards, using explicit criteria, and all found that then-current grading practices typically did not do that. Standards-based grading reforms, beginning in the 1990s and gaining more traction in the 2000s, are intended to remedy this situation. The movement is gaining momentum, but definitive studies are not yet available.

Much research on teachers' grading has used a practical, rather than theory-based, approach and simple survey methods. General findings include the following (Brookhart, 1994, 2013): Teachers try hard to be fair to students, including informing them of what will contribute to their grade. Achievement measures, especially tests, are the major components in grades; teachers commonly also consider effort and ability. Elementary teachers use more informal evidence and observation than secondary teachers, who base grades more on paper-and-pencil tests and other written activities.

Finally, teacher grading practices vary widely, partly because different teachers view the meaning and purpose of grades differently (Frary et al., 1993; Vanlommel & Schildkamp, 2019).

Two recent research agendas (McMillan, 2001; McMillan et al., 2002; Randall & Engelhard, 2009, 2010) support the conclusion that while teacher grading practices are much maligned, in fact academic achievement is the factor teachers consider most heavily. Teachers reported considering academic enablers (McMillan, 2001), such as effort, in students' grades as a way to evaluate students' academic engagement, which some teachers value as part of school learning. Randall and Engelhard (2010) also found that academic achievement is the factor teachers primarily consider when assigning academic grades. They found interactions among ability, achievement, behavior, and effort: with high effort and behavior, low-achieving, low-ability students receive an average grade of C+, and in general, high effort and behavior boosts the grades of students of any ability somewhat. Thus, studies continue to find that teachers' grading practices mix achievement and nonachievement factors, but it seems clear now that achievement is the main component of grades, and teachers' use of other factors is generally reported to be motivated by what in their view is good for students.

If, in general, improving teacher assessment literacy is difficult (C. Campbell, 2013), in grading it is even more difficult because of the multiple conflicting viewpoints teachers hold about grades. For example, Olsen and Buchanan (2019) investigated changes in the understanding and grading practices of teachers involved in year-long professional development designed to reform grading practices in two secondary schools. Change did occur, but it was partial and did not necessarily have the intended effects. Teachers sometimes adapted recommended strategies and then refined them to fit their classroom context, in the process using a belief system that the grading reforms were meant to change.

Students, too, need some basic assessment literacy to understand and benefit from the feedback inherent in the grades they receive. Almost by definition, successful students navigate this well, eagerly wringing any information they can get from their grades to learn more and continue to do well in school (Brookhart, 2001). Students need to learn that grades apply to their learning and not to them as people, that grades are changeable, and that, at least to a large degree, that change is under their control. These lessons, of course, only hold true in situations where teachers' grading practices are supportive of learning.

IMPACT OF GRADING ON LEARNING Grades constitute a form of summative feedback that is used for reporting and certifying student learning. However, successful students also use grades as part of their self-regulation of learning, as they learn to become self-monitoring; as part of their self-efficacy development, as they learn how grades sum up their performance; and as part of their metacognition, as they use grades as one piece of evidence to judge the quality of their own understanding (Brookhart, 2001).

Students consider grades a type of feedback. For example, Harris et al. (2014, p. 121) asked a sample of New Zealand upper primary and lower secondary school students to create a free-response drawing of their experience of feedback—44% of the drawings included grades or scores.

Lipnevich and Smith (2009a,b) pointed out that the most common type of feedback students receive in class is grades. Nevertheless, the U.S. college students in their focus groups deemed grades unnecessary if the purpose of the activity was to learn. In addition, they exhibited negative emotions when receiving a low grade; some became angry at the instructor and others reported dissatisfaction with themselves or embarrassment. In a survey of teacher education students in Australia, “stated grade” was rated as the least useful type of feedback with the exception of group verbal feedback; “written summary” and “brief comment throughout” were rated most useful (Ferguson, 2011).

College students want feedback even on graded work (Pitt & Norton, 2017), using the term *feedback* to mean comments. This study of UK university students showed they interpreted grades and any additional feedback differently depending on their feelings of competence, their feelings about their lecturer, and their level of emotional maturity. One of the students in Pokorny and Pickford’s (2010) focus groups of college students in the United Kingdom commented that feedback (comments) accompanying final graded work was moot, since “it’s way too late to do anything about it” (p. 24).

The effects of grading practices on students’ motivation to learn has also been studied. T. J. Crooks (1988) and Covington (1992) both showed the effects of school evaluation practices on students. Covington (1992) called for practices that fostered what he called “motivational equity” (p. 21); not all students can learn the same things, but all students should experience evaluation practices that support student learning (p. 21):

Everyone can experience feelings of resolve and a commitment to think more, and to dare more; feelings of being caught up in the drama of problem solving, and of being poised to learn and ready to take the next step.

He showed that low grades do not accomplish this. Lohbeck (2019) found that for a sample of German fourth graders, self-reported grades in mathematics and German predicted academic self-concept in the respective domain (but not the other), which in turn predicted self-perception of effort in the domain. In other words, students whose grades give them evidence that they are effective learners in a domain may be motivated to continue to expend effort in learning in that domain.

EVALUATION OF GRADING INFORMING TEACHING AND LEARNING Grades are used for wide range of purposes, some of which they serve better than others. Grades are multidimensional measures of situated school learning and correlate only moderately with standardized achievement measures. Grades have a long history as assessment

information used to inform teaching and learning. There is evidence that grades can be supportive of student learning if certain grading practices are followed, although there is great variation in grading practices. Grading may be one of the most difficult areas in which to support and improve teacher assessment literacy because many people (teachers, parents, students, etc.) have strongly held opinions and feelings about grading, based on their own grading histories.

State Summative Assessment

High-stakes, large-scale assessment has affected teachers' instructional practices (Faxon-Mills et al., 2013; Pedulla et al., 2003); that is, state summative assessment has been and is being used to inform teaching and learning. However, the ability of such assessments to inform teaching and learning is limited, at best.

PURPOSE AND USES OF STATE SUMMATIVE ASSESSMENT State summative assessments are typically administered beginning in March or April and are not intended to provide information to inform the instruction of students within the current academic year. Rather, state summative assessments, also referred to as state accountability tests or simply state tests, are intended to serve state and federal accountability purposes, including district-, school-, teacher-, and student-level accountability. State summative assessments may provide information that is useful to teachers to evaluate their instructional practices or useful to schools and districts to evaluate their curricular and instructional programs.

DESIGN OF STATE SUMMATIVE ASSESSMENT Beginning in the 1990s, many state assessments were designed to serve as examples of the type of assessment that should be occurring in the classroom. The increased use of constructed-response items, direct writing prompts, and performance tasks within traditional end-of-year assessments and alternative formats such as portfolio assessments is evidence of this desire to give students a test worth teaching to (Resnick & Resnick, 1992; Rothman, 1995; Shepard, 1991). For example, one state wrote that the purpose of state assessments was “to improve classroom instruction by (a) providing useful feedback about the quality of instruction and (b) modeling effective assessment approaches that can be used in the classroom” (Massachusetts Department of Education, 1999, p. 4).

The implementation of NCLB in the early 2000s brought both increased stakes in terms of school accountability and increased testing with the requirement of annual testing of all students in Grades 3 through 8 plus one grade in high school (NCLB, 2001). The 2010s further increased the stakes and footprint of state summative assessments in the classroom with longer tests, more complex standards and tasks, the declaration of an “honesty gap” between national test scores, state test scores and teacher grades, and the use of test results for teacher evaluation (Achieve, 2013). These factors converged to create an expectation that state summative assessments can, should, and must provide “actionable information” to inform teaching and learning (CCSSO, 2013;

U.S. DOE, 2015), another instance of the pressure to extract such information even from assessments not designed for it.

The primary information from state summative assessments is an indication of student performance in relation to state achievement standards in terms of scale scores and achievement levels. State summative assessment operationalizes the state's content standards and performance expectations, showing teachers how the state is interpreting the content standards. In many state programs, this information is supplemented by sets of released items with annotated samples of exemplar student responses to constructed-response items or essays. All of this information is designed to help increase teachers' and administrators' understanding of the state's interpretation of the content standards and expectations for students at each grade level. Teachers and administrators are expected to use such information to calibrate their expectations for student performance with those of the state (Madaus et al., 2009; Parsi & Darling-Hammond, 2015).

MEASUREMENT CONSIDERATIONS IN STATE SUMMATIVE ASSESSMENT Federal peer review has ensured that the process of gathering data and producing scores on state summative assessments is typically sound. Operational best testing practices employed routinely by assessment contractors in conjunction with federal requirements and peer review for state assessments systems, including requirements regarding accessibility and the availability of accommodations, and the commonplace practice of convening technical advisory committees for state testing programs help ensure the quality of the process and the validity of the results, although the availability and quality of validation studies related to state assessment programs remains a concern. State assessment programs, however, are not immune from real and perceived errors or irregularities in test administration, scoring, or reporting, which are often consequential and highly publicized (Rhoades & Madaus, 2003). In addition, there are long-standing concerns about the generalizability of results from state summative assessments and test score inflation due to test preparation and a narrowing of instruction to the particular standards, or portions of standards, included on the state summative assessment (Koretz, 2017).

The process of designing state summative assessment programs for school accountability, including the selection (or rejection) of items based on their statistical properties, however, limits the usefulness of the information gathered through state summative assessments to inform teaching and learning because the assessments' design focuses on the summative purpose. State assessments are designed to produce information about overall student achievement of, or proficiency on, state standards, focused appropriately on the validity of inferences related to these broad student outcomes. In contrast, most instructional decisions require more fine-grained diagnostic information. In addition, measures to inform teaching and learning need to provide timely information about student performance and the instructional process, and they should be based on an explicit theory of student learning and a theory of action related to improving instruction (Ing et al., 2021).

Measures that inform improvement in instruction need to be explicitly linked to a theory of improving instruction (Bryk et al., 2015). In contrast to measures for the purposes of accountability, measures for improvement provide timely information about the improvement process to address questions such as, How can processes central to the theory of improvement be altered to be improved? (Solberg et al., 1997). That is, rather than indicating whether a student has learned something, measures for improvement speak to the processes that support student learning instead of focusing solely on the outcomes of student learning. Not surprisingly, multiple measures are needed for this ambitious purpose of improving instruction at scale (see, e.g., Takahashi et al., 2022). No single measure can provide all the necessary information for what is required to increase the quality of instruction (Bennett, 2015; Bennett & Gitomer, 2009; Newton, 2010). Instead, a comprehensive system of measures is needed to gather the necessary information to inform instructional improvement efforts, including outcome measures such as indicators of student learning (ranging from daily analysis of students to student performance on annual standardized achievement measures) as well as measures for improvement that inform the daily work of those in the system who are working toward the same improvement goals. In such situations, the same measures may be used for different purposes by different users (Gitomer & Duschl, 2007).

Increased pressure to provide more information and more actionable information from lengthy state summative assessments has led to the reporting of subscores for individuals and groups of students. Given that most state summative assessment programs are designed to fit unidimensional item response theory models and often contain a small number of items measuring specific content standards or strands, subscores provide little, if any, useful, reliable information to inform teaching and learning (see Zenisky et al., this volume).

IMPLEMENTATION OF STATE SUMMATIVE ASSESSMENT IN PRACTICE Expectations that state assessment and accountability data could or should affect teaching and learning in some way still exist. That rhetoric abounds on state and district education websites. The U.S. DOE has encouraged the use of district data systems, including state summative test results, for informing teaching and learning and suggested data literacy supports for teachers (U.S. DOE, 2009, 2011). At the beginning of the NCLB era, there was concern about state test data having negative effects on instruction. For example, Pedulla et al. (2003) classified states according to the severity of the state summative test's consequences for districts, schools, and teachers and for students. They surveyed classroom teachers and found that teacher responses varied with the severity of stakes and with level (elementary/middle/high school). Teachers reported feeling pressure to raise scores on the state-mandated tests. A large percentage of teachers reported pressure to deliver instruction that ran counter to their own beliefs about good teaching practice.

There is some evidence that some teachers use state assessment results to inform changes in curriculum content and emphasis, changes in pedagogy, and changes in teacher interactions with individual students (Faxon-Mills et al., 2013). Changes in curriculum included changes in sequence and emphasis of topics, a focus on basic skills, and a focus on higher order skills, depending on the state and study. Changes in pedagogy included a focus on test preparation and changes in instruction and assessment practices. Changes in interactions with students included individualizing instruction and focusing on students who were close to the proficiency cutoff and could potentially cross it and increase pass rates; these students have been referred to as “bubble kids” (Faxon-Mills et al., 2013), an unintended consequence of NCLB accountability requirements that has been alleviated by the use of indicators that consider student performance across the achievement continuum. Similarly, Au (2007) found that the primary effects of high-stakes testing on teaching were curriculum narrowing, fragmenting subject area knowledge into test-related pieces, and increased use of teacher-centered instruction; however, in some cases the high-stakes tests led to curriculum expansion, synthesizing subject area knowledge, and student-centered instruction. Some of these changes might have potentially positive effects on learning, for example, focusing on student thinking or integrating knowledge, and some might have negative effects, for example, focusing on some students more than others or increasing lectures and other teacher-centered instructional methods.

More recently, some studies have reported that while some teachers use state test results, at least in a general way, classroom assessment—both summative and formative—is more closely linked to decisions about changes in instruction. For example, Wilkerson et al. (2021) found that Nebraska teachers’ use of state summative, interim, and formative assessment data to inform instruction was consistent with the frequency of administration of each type of assessment, with data from state summative tests being used least often.

Farrell and Marsh (2016) studied five high-needs middle schools in three districts. They found that state assessment results were used mostly at the beginning of the school year for grouping students but were not associated with other changes in instruction.

IMPACT OF STATE SUMMATIVE ASSESSMENT ON LEARNING Any impact of state summative assessment on student learning would be mediated through its impact on curriculum and teaching, as described in the previous section. Given this distal connection to teaching and learning and its intended purpose as a summative tool for accountability, one might argue that it is inappropriate to even ask questions about the impact of state summative assessment on learning. A primary purpose of federally mandated state assessment programs, however, is to support Title I, which is intended “to provide all children significant opportunity to receive a fair, equitable, and high-quality education, and to close educational achievement gaps” (Every Student Succeeds Act, PL 114-95, 2015, Sec. 1001. Statement of Purpose).

Therefore, one way to examine the impact of state summative assessment on student learning that might be appropriate is in the way that it was intended to be used: that is, in terms of the extent to which test scores indicate that progress has been made in providing equity in educational opportunity and closing achievement gaps. Results from NAEP and international assessments indicate that the achievement gap by socioeconomic status has remained large and stable since the 1970s (Hanushek et al., 2019). Results from the NAEP Reading and Mathematics tests at Grades 4 and 8 since 1992 showed modest impact on the pace of reductions in the gap in performance between Black students and White students through 2005 (Braun et al., 2010). NAEP results showed little change in achievement and in the size of the White–Black and White–Hispanic achievement gaps from 2010 to 2020 (U.S. DOE et al., 2020).

EVALUATION OF STATE SUMMATIVE ASSESSMENT INFORMING TEACHING AND LEARNING State summative assessment programs are designed to provide summative information for accountability and administrative purposes. The growing clamor for more formative information from all assessments perhaps causes more discomfort in the area of state summative assessment than for other types of assessment because its intended purpose and design are decidedly summative. It remains to be seen whether state summative assessments can be redesigned to better provide desired formative information.

Some states are considering the use of through-year assessment models, which emerged as an alternative to a single end-of-year test during the 2009–2010 Race to the Top assessment competition, in an attempt to provide both formative (i.e., timely and instructionally useful) information and the required summative information from a single assessment program. To date, many of those efforts have been limited to replacing the single end-of-year state summative test with the traditional interim assessment model in which all standards are assessed on a short test administered three or four times per year, but some states are attempting to better synchronize the sequence of standards tested throughout the year with instruction or to closely connect the state assessment with the taught curriculum (Powell et al., 2022).

Using information from a single assessment event for both formative and summative purposes poses a significant challenge to the interpretation and use of through-year assessment models for summative accountability purposes (Dadey et al., 2023; Gong, 2021). Some states and state assessment programs, such as Smarter Balanced Assessment Consortium, have responded to this not by changing the summative assessment, but by offering optional interim assessments of varying lengths (full comprehensive assessments, interim assessment blocks, or focused interim assessment blocks, depending on how many standards are assessed) aligned to the state summative assessment. Teachers can choose to administer and use these interims for formative purposes and even create their own from item banks. In other words, Smarter Balanced has responded to this pressure on summative assessment by shifting resources to interim assessment (Smarter Balanced, 2019).

ASSESSING LEARNING TRAJECTORIES FROM CLASSROOM LESSONS TO SUMMATIVE

The assessments described in the previous section were measures of performance or learning status at single points in time, and this evidence about achievement status was what was leveraged to inform teaching and learning. In contrast, the assessments described in this section offer information for informing teaching and learning by looking at student achievement trajectories. This work represents a distinction from assessments in the previous section in two ways: It looks at performance over time and it is theory based—in contrast with any growth modeling using the more traditional measures.

Learning implies growth or change. Growth or change can be measured or indicated in several different ways: as trend lines, pre–post (or pre–mid–post) designs, or progress on a map of development in a domain of learning. Different assessment methods have been applied to show learning spans or trajectories, and three of these are described in this section. In this chapter, the generic term *assessment of learning spans or trajectories* is meant to include any assessment method based on trajectories. The term *learning progressions* has come to be applied to one particular method, described below in the section “Assessment Based on Learning Progressions.” However, the idea of learners progressing from naive to expert, or from novice to mastery, in a particular domain of content underlies all three of the assessment methods in this section. Assessments of trajectories spanning from classroom lessons through summative (outcome) achievement rely on individual measures taken over time. Inferences are then made from students’ trends or growth, as well as the final outcome or level achieved, and this trend information is used to inform teaching and learning.

Assessing trajectories of student achievement fits with the formative assessment cycle as described by the question loop in Table 15.1: Where are we going? Where are we now? How will we get there? Assessments that focus on learning trajectories ask those same questions, but in terms of a bigger developmental picture than when those questions are used for classroom formative assessment of small sequences of lessons. In both, the emphasis is on the formative, on “getting somewhere.” Unlike state standards or classroom learning goals, learning trajectories describe both the destination and the way to get there. This makes learning trajectories potentially a rich source of formative information.

This section considers three assessment methods focused on learning spans or trajectories. Curriculum-based measurement (CBM) has the longest history of the three, in both research and practice. Student learning objectives (SLOs) date from the recent standards-based accountability era and arose in a practical context. Assessments based on learning progressions seek to connect theory about how students learn various topics or concepts with their assessment. Learning progressions are the subject of several ongoing research agendas.

Curriculum-Based Measurement

CBM is an approach to measuring the academic growth of individual students during instruction toward a general outcome of interest (Deno, 2003a). The measurement process is based within the curriculum, and the measures themselves are carefully constructed measures of that general outcome (i.e., the endpoint), rather than being drawn from or based on a particular curriculum. Repeated measurements of student achievement of the general outcome provide information to the teacher regarding a student's progress toward the ultimate learning goal, signaling whether changes to instruction are needed.

Figure 15.1 shows a sample graph of CBM measures collected over 27 weeks of instruction, with an instructional change occurring after Week 16 (Espin et al., 2018). Key components of the CBM progress-monitoring graph include: (a) baseline data, representing the student's current level of performance; (b) peer data, representing typical performance and reflecting the discrepancy between the student and peers; (c) a goal line, representing the expected rate of growth and end-of-year level of performance; (d) data points, representing the number of correct and incorrect responses on weekly probes; (e) slope or growth lines, representing the student's rate of growth over time; and (f) solid vertical lines, representing instructional changes. The slope of the student's performance during a particular instructional intervention provides feedback to the teacher of the success of that instructional approach for the student.

Purpose and Uses of CBM

CBM was developed in the 1980s to support the instruction of individual students with disabilities (Deno, 1985, 2003b). CBM was developed as a means “for teachers to use technically sound, but simple, data in a meaningful fashion to document student growth and determine the necessity for modifying instructional programs” (Stecker et al., 2005, p. 795).

Since its initial development as a tool for elementary school teachers of students with disabilities, the applications and use of CBM and progress monitoring have expanded to other grade levels and multiple content areas, general education, and other uses, such as program evaluation (Allen & Smith, 2022; Jenkins & Fuchs, 2012). A major application of the principles of CBM has been as part of the large-scale Response-to-Intervention program (Fuchs & Fuchs, 2007; National Center on Response to Intervention, 2010). The use of CBM for program evaluation and large-scale screening programs such as Response to Intervention is not directly related to the use of assessment information collected during instruction to inform instructional decisions and therefore is not a focus of the discussion of CBM in this chapter.

Design of CBM

From its inception, CBM has focused on the frequent collection of student performance on a general outcome measure linked to students' overall proficiency on long-term goals (Deno, 1985, 2003a; Stecker et al., 2005). The purpose for using the general outcome measure, rather than more discrete measures tied to individual standards or particular

knowledge and skills, was to provide teachers with clear evidence of student progress toward the desired goal, that is, the level of proficiency desired at the end of the year or other defined instructional period (Fuchs & Deno, 1991). The goals are specific to the content area assessed and the grade level of the students. The general outcome measures are intended to focus on those academic performances that represent “vital signs of educational development” (Deno & Mirkin, 1977, p. 14).

The vital signs measured on individual assessments used in CBM are to be regarded as indicators of student performance within a domain, rather than as direct measures of specific knowledge and skills within the domain, in some ways consistent conceptually with the manner in which state summative tests are designed to provide an indicator of students’ overall proficiency and not a measure of student mastery of individual standards (Fuchs, 2016). Central to the vital signs approach to measurement in CBM is the belief that (a) teachers, in many cases, have no simple indicators with which to monitor the effects of instruction and (b) information that teachers are more likely to have, student mastery of individual standards or skills, is not necessarily evidence of increasing student proficiency on the general outcome (Deno, 1993).

A progress-monitoring graph depicting student performance over time, like the one in Figure 15.1, is one of the fundamental components of CBM (Espin et al., 2018).

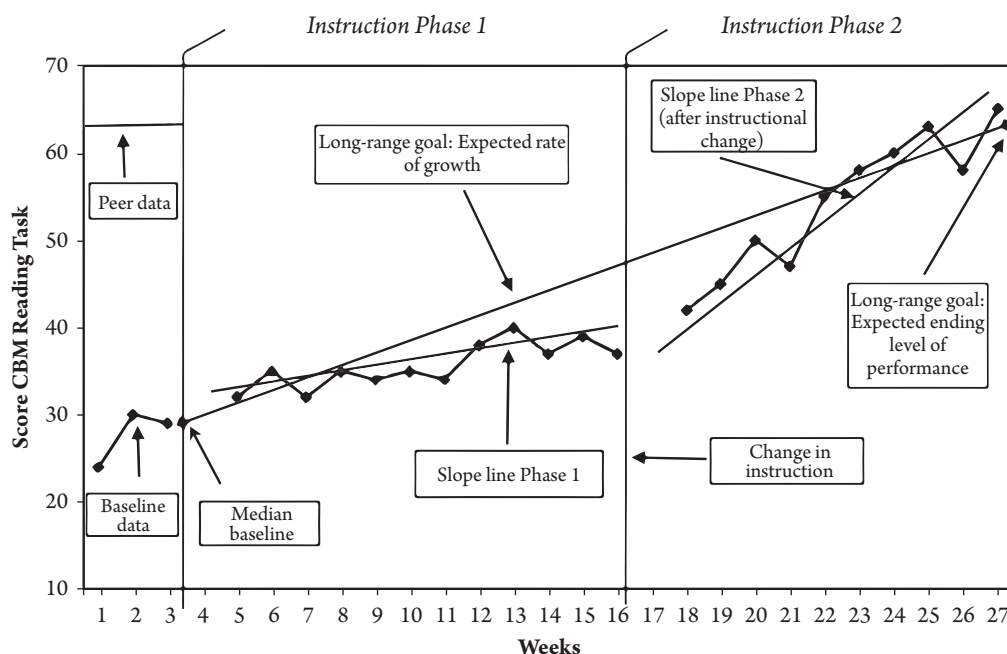


FIGURE 15.1

Sample Graph of Curriculum-Based Measurement (CBM) Measures Collected Over 27 Weeks of Instruction, With an Instructional Change Occurring After Week 16

Note. From “Curriculum-Based Measurement Progress Data: Effects of Graph Pattern on Ease of Interpretation,” by C. A. Espin, N. Saab, R. Pat-El, P. D. M. Boender, and J. van der Veen, 2018, *Z Erziehungswiss*, 21(4), 767–792. <https://doi.org/10.1007/s11618-018-0836-9>. Copyright Creative Commons Attribution 4.0 International License. Reprinted with permission. (<http://creativecommons.org/licenses/by/4.0/>).

To evaluate the effectiveness of instruction for a particular student, the teacher examines the graph to determine whether the student is progressing at the desired rate of growth and whether the student will achieve the goal (van den Bosch et al., 2017).

By comparing the slope of the line depicting student progress to the line representing anticipated growth over time toward the long-term target, teachers can determine whether an instructional change is warranted (Stecker et al., 2005). Consistent below-target performance is regarded as evidence that a change in instructional program might be needed. Consistent above-target performance may suggest that the teacher consider raising the long-term goal. In addition to information about an individual student's progress toward the long-term goal, teachers may also consider normative information (e.g., student progress relative to others in the class) or group-level information (e.g., median student progress in a class).

Measurement Considerations in CBM

The quality of CBM information is enhanced by the use of technically sound assessment instruments developed for particular grade levels, content areas, and skills (Stecker et al., 2005; Marston, 2012). Each assessment probe is developed to be a measure of the same overall performance and equal in difficulty to all other probes measuring that area of proficiency. A critical component of CBM, therefore, is the selection or development of multiple alternate forms of an assessment instrument, each of which is a sufficiently reliable indicator and supports valid inferences of students' overall proficiency. Consequently, true applications of CBM cannot rely on teacher-made assessments unique to an individual teacher or group of teachers or on the ad hoc selection of commercially developed assessment instruments. CBM requires the use of assessments that have been empirically selected to meet the technical requirements outlined in this paragraph (Deno, 2003a,b).

An extensive body of research describes the type and technical quality of measures used in various content areas: reading (Fuchs, 2004; Miura Wayman et al., 2007; Tindal, 2013); writing (McMaster & Espin, 2007; McMaster et al., 2011), and mathematics (Foegen, 2008; Fuchs, Fuchs, Compton, et al., 2007; Fuchs, Fuchs, & Hollenbeck, 2007; Gersten et al., 2011). Research also describes measures used with various groups of students, for example, students at secondary levels (Espin et al., 2018; J. Shin & McMaster, 2019), English learners (H. Campbell et al., 2013), and students who are deaf or hard of hearing (Lam et al., 2020).

Fuchs (2004, p. 189) described two approaches to the task of identifying "measurement tasks that simultaneously integrate the various skills required for year-end performance" as (a) identifying tasks that "correlate robustly (and better than potentially competing tasks) with the various component skills constituting the academic domain" and (b) a "systematic sampling of the skills constituting the annual curriculum to ensure that each weekly CBM represents the curriculum equivalently." Tasks at early elementary grades

are more likely to reflect the first approach and tasks at upper elementary and secondary grade levels are more likely to reflect the second approach (Allen & Smith, 2022).

Although linked directly to students' overall proficiency on the desired construct, the vital signs approach of CBM relies heavily on the use of the predictive validity of the measures and has given less weight to issues related to alignment of the individual measures to specific content standards (Deno, 2013), raising concerns and criticisms about the appropriateness of the measures. CBM assessment probes used at higher grade levels in general education classrooms and in content areas other than reading have focused more on alignment to content standards (Allen & Smith, 2022; Fuchs, 2004; Morton, 2013). Recent research has focused on the alignment of CBM measures to state content standards and has used performance on state assessments as a criterion for evaluating the technical adequacy of CBM measures (Shriner & Thurlow, 2012; Tindal, 2013).

Implementation of CBM in Practice

After the selection of assessments, the key steps in the CBM process are regular administration of the assessments, scoring and graphical depiction of student performance, and appropriate interpretation of student progress toward the long-term goal. From its beginnings, an important feature of CBM was the use of assessment instruments that were short, easy to administer, and easy for teachers to score. Such features were intended not only to increase the reliability of scores, but also to increase the utility of the process; that is, such features were considered critical to enhancing the likelihood that teachers would administer the assessments and collect data on student performance on a regular basis.

CBM arose from a different view of teaching and learning than classroom formative assessment, one more focused on affecting teacher instructional adjustments and thereby enhancing student achievement (Stecker et al., 2005), similar in focus to formative evaluation for mastery learning (Bloom et al., 1971) rather than informing students' regulation of learning directly. Although described as a tool to support data-based or data-driven decision-making (Fuchs & Deno, 1991; Hosp & Hosp, 2012) the development and implementation of CBM is not a product of the accountability-centered data-driven decision-making models described as a policy theory of action (Shepard et al., 2017). Rather, with its foundations in special education and the instruction of individual students by special education resource teachers, CBM is more aptly classified as an application of single-subject research design or time series design (Marston, 2012).

In contrast to the data-driven decision-making models assumption that teachers will know how to help students (Shepard et al., 2017), CBM is based on the assumption that a teacher does not know in advance whether, or how well, a particular instructional intervention will work for an individual student. In their Data-Based Program Modification manual, a precursor to CBM, Deno and Mirkin (1977, p. 22) offered the following as two fundamental assumptions of Data-Based Program Modification:

Assumption 1: At the present time we are unable to prescribe specific and effective changes in instruction for individual pupils with certainty. Therefore, changes in instructional programs which are arranged for an individual child can be treated only as hypotheses which must be empirically tested before a decision can be made on whether they are effective for that child.

Assumption 2: Time series research designs are uniquely appropriate for testing instructional reforms (hypotheses) which are intended to improve individual performance.

The theory of action for CBM, therefore, can be described as experimental teaching or data-based instruction in which the teacher takes an active role in verifying that an instructional strategy is effective for the individual student (Fuchs & Bradley, 2012). The fundamental assumption of CBM is

successful intervention required that teachers receive clear and unambiguous feedback regarding the general effects of their instructional efforts. If teachers are either uncertain about the overall effects of their efforts, or believe they have been successful simply because a student learns the specific content that has been taught, their efforts to improve growth will be unsuccessful. (Deno, 2003a, p. 6)

In early applications of CBM, individual teachers were required to create their own progress-monitoring graphs, but with the assistance of computer-based applications to generate the graphs, the focus of teacher assessment literacy needs has shifted to the ability to read and interpret the information provided by the graphical depiction of student progress (Fuchs & Fuchs, 2007). This requires a basic understanding of how to interpret fluctuations in student performance over shorter time intervals in relation to changes in performance over larger time intervals (van den Bosch et al., 2017). With training, teachers, on average, become fairly proficient in reading the data on CBM graphs, but there is variation in the ability of individual teachers (Espin et al., 2017). Further, teachers have greater difficulty in interpreting the data on the graphs, particularly ambiguous data, and linking it to instructional decisions (van den Bosch et al., 2017). Although studies have shown that students make more progress in achievement when teachers use graphs than when they simply record data (Fuchs & Fuchs, 1986), the recent research on teachers' ability to read and interpret graphs highlights the need for professional development in interpreting CBM data and using it in conjunction with other data about instruction and student performance to inform instructional decisions. A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making (Gesel et al., 2021) found that professional development supports could have a positive impact on teachers' data-based decision-making skills.

Impact of CBM on Learning

Relatively few studies have directly examined the impact of the use of CBM on student learning (Deno, 1985). Stecker et al. (2005, p. 795) stated, "The hope was that by responding instructionally to students' poor patterns of performance, teachers should

be able to enhance student achievement.” Two studies that synthesize the research studies that have been conducted in this area indicate that the proper use of CBM by teachers as part of instruction can have a positive impact on student learning (Jung et al., 2018; Stecker et al., 2005). Both studies, however, describe significant challenges to teachers implementing CBM with fidelity and using the results of measurement to inform modifications to instruction.

Stecker et al. (2005) described research with students with disabilities from the initial development of CBM in the 1970s and research conducted with students in general education settings that began in the 1990s. Critical variables examined in the studies reviewed include the use of data-based decision rules, the inclusion of a skills-analysis component to the reporting of CBM results, and the provision of recommendations for next steps by teachers. Across the studies for students with disabilities and in general education settings, Stecker et al. (2005) found that the use of CBM alone did not have a significant positive impact on student learning (i.e., progress monitoring without making use of the data), but that significant gains in student achievement were attained under the following five conditions: (a) teachers respond to the CBM data to tailor instruction to student needs; (b) a preestablished data-decision framework is adhered to; (c) computer applications are used to improve the efficiency of all aspects of the CBM process; (d) skills analysis, in conjunction with consultation, is provided; and (e) some form of recommendations for next steps are provided. Studies conducted in general education settings also demonstrated that instructional modifications based on class-wide CBM data, rather than data for individual students, could be used to produce gains in student achievement and that with training, CBM data could be used successfully by students in peer-tutoring contexts to improve student achievement.

Jung et al. (2018) conducted a meta-analysis of the effects of CBM, under the generic label of data-based individualization (DBI), on the performance of students with intensive learning needs. They examined the effect of teachers’ use of DBI on student achievement and the factors that influence the strength of effects of DBI. Of the 14 studies included, 6 were conducted after the original Stecker et al. (2005) review. Overall, Jung et al. (2018) found that the use of DBI alone and the use of DBI with additional information or supports, DBI Plus, had significant positive effects on student achievement compared to a “business-as-usual” control. Consistent with the findings of the earlier review, they found that providing teachers with some type of support for the interpretation and use of the progress monitoring data to make instructional modifications had a significant impact on the efficacy of using DBI to improve student achievement.

The findings from both reviews on the impact of CBM on student learning is consistent with the definition of formative assessment adopted by the authors for this chapter, that is, that a practice is formative to the extent that the evidence about student achievement is not only elicited, but also interpreted and used by teachers and students to make decisions about instruction. Research on the impact of CBM on student learning

is also consistent with the findings elsewhere in this chapter that additional research is needed on teacher assessment literacy, and additional ongoing support is needed for teachers' use of data to inform instructional decisions.

Evaluation of CBM Informing Teaching and Learning

CBM is supported by an extensive body of research compiled over more than 3 decades since its inception (Allen & Smith, 2022; Espin & Wallace, 2004; Stecker et al., 2005; Tindal, 2013). Much of the research on CBM can be classified into three categories that are critical to developing a validity argument for its use to inform teaching and learning (Fuchs, 2004): (a) research on the technical adequacy of the measures and scores produced from each assessment, (b) research on the technical features of the slope (i.e., student performance over time) to determine whether increasing CBM scores are in fact associated with improvement in overall competence in the academic domain, and (c) studies concerning instructional utility to determine whether practitioners can use the CBM information to improve instructional decisions and student achievement. This third category of research is particularly relevant to the use of CBM to inform teaching and learning.

Tindal (2013) described the dual focus of the CBM research as grounded in a “measurement paradigm” and focused on a “training in data use and decision-making paradigm”; although the former is an essential foundation for the use of CBM to inform teaching and learning, the latter is critical to ensuring that it can and will be used appropriately to inform teaching and learning. As with formative assessment processes, a long-acknowledged threat to the use of CBM to improve teaching and learning is that when the data indicate that a change in instructional intervention is needed, teachers do not know how or what to adjust in their instruction (Fuchs et al., 1989; Stecker et al., 2005). Recent research has identified teachers' aforementioned difficulty in reading and interpreting CBM graphs (Espin et al., 2017; van den Bosch et al., 2017) and teachers' insufficient professional development in making instructional decisions based on data as two additional threats to the use of CBM to inform teaching and learning (Espin et al., 2021).

Student Learning Objectives

SLOs arose at the local level in the context of teacher evaluation policy, giving teachers some say in the measures for which they would be held accountable. The locus of SLO shifted to the state level when teacher evaluation based on teachers' impact on student learning became a federal requirement for states receiving Race to the Top funding or NCLB waivers. Using SLOs meant teachers of subjects that were not represented in state accountability tests could be held accountable for student progress in their subjects. This brief section on SLOs is included in the chapter for the sake of completeness because the intent of SLOs is that student learning will be affected. However, there is little evidence to date that student learning is in fact affected.

Purpose and Uses of SLOs

SLOs—also referred to as student learning goals, student learning outcomes, student learning measures, and student growth objectives—delineate a process designed to integrate assessment and instruction for the purpose of improving student learning. Typically, teachers and building principals together negotiate the measures to be used for the SLO process in teacher evaluation (New York State Education Department, 2014; Texas Education Agency, 2021). The SLO process is intended to include (a) developing carefully planned learning goals for what students will learn over a given period of time and (b) evoking critical, evidence-based thought about student progress and growth through instruction (Community Training and Assistance Center [CTAC], 2018).

Initially used as part of innovative pay-for-performance compensation programs (CTAC, 2004, 2013), SLOs were developed as a school improvement tool to be implemented collaboratively by administrators and teachers at the local district and school levels (Diaz-Bilello & Thompson, 2019). SLOs gained popularity as they were adopted by many states as the primary method for evaluating teachers' impact on student learning in content areas and grade levels in which there was not a state assessment (Marion & Buckley, 2012) in response to federal requirements under Race to the Top and NCLB waivers (Briggs et al., 2019).

Design of SLOs

The use of the SLO process as a tool for school improvement has been described as a school-based application of the management-by-objectives model (Chapman, 2014). The design and implementation of individual SLOs by teachers under the CTAC (2013) model is driven by six elements: (a) learning content (what will be learned), (b) instructional strategies (approach to teaching), (c) interval of instruction (time frame for teaching), (d) assessment (measures of learning), (e) student growth targets (numeric goals for learning), and (f) student population (who is included).

Those elements are then implemented in a cyclical process that includes periods of instruction, assessment, and adjustment. A common SLO cycle might include the following steps (Diaz-Bilello et al., 2016; Diaz-Bilello & Thompson, 2019): identify the learning goal, develop or select assessments, analyze baseline data, establish targets, analyze a cycle of instruction and assessment, revise targets when appropriate, analyze a cycle of instruction and assessment, and complete the SLO evaluation. Each cycle of instruction would also include the use of formative assessment procedures and adjustments to instructional strategies, as needed.

Measurement Considerations for SLOs

A critical distinction between SLOs designed for accountability and SLOs designed for improving instruction is the focus on comparability across districts, schools, and

teachers, as evidenced by the level of autonomy allowed in two critical aspects of the SLO process: the development and selection of assessments and the establishment of growth targets (Cushing & Meyer, 2014; Diaz-Bilello & Thompson, 2019). It is less important for SLOs designed to improve teaching and learning than for SLOs for teacher evaluation to be comparable across contexts (Lane & DePascale, 2016). Even when SLOs are used for teacher evaluation, however, there is a large degree of variation across states in how SLO performance targets have been established (Crouse et al., 2016).

Using SLOs requires a basic understanding of reliability of observed scores, in general, and difference scores, in particular. Decisions made about individual students' meeting their growth target will be affected by error in the pretest and posttest scores. For example, a common approach to setting growth targets for SLOs is referred to as the "half-the-distance" method, in which the growth target on a posttest is defined as the score that is halfway between the student's baseline score on a pretest and either a specified criterion score or the maximum score on the posttest (i.e., a perfect score; Austin Independent School District, 2015; Missouri Department of Education, 2015). The use of approaches such as half-the-distance can produce different results based on the level of student performance and the design and type of assessments used (Austin Independent School District, 2015) and can produce results that are not acceptable within the given context, such as target scores that are below the accepted passing score on the posttest (English et al., 2015).

Implementation of SLOs in Practice

When implemented with fidelity, the SLO process should provide feedback to teachers, administrators, and students about the specific learning content and growth targets expected to be achieved during the interval of instruction. It also should provide feedback to teachers and administrators about the instructional strategies teachers will use to support student learning and attainment of growth targets. At the end of the SLO process, teachers and administrators receive summative feedback on whether a SLO or multiple SLOs implemented by individual teachers have been met. That decision is usually based on the percentage of students included in the SLO process who met their individual student goal (Hall et al., 2014).

The SLO process requires a high level of training and support at the district and school levels and is difficult to implement without an extended commitment to the process (Schneider & Johnson, 2018). In addition to general assessment literacy, specific teacher needs include the ability to clearly specify measurable objectives that are also primary curriculum goals and to select or create assessment tools whose results are valid, reliable, and fair indicators of those goals. Teachers and administrators need the ability to analyze and interpret assessment results against both student learning goals and teacher professional goals for purposes of high-stakes decision-making.

Impact of SLOs on Learning

Few research studies have been conducted examining the extent to which the use of SLOs provides feedback that teachers use to inform their instructional strategies and decisions during the implementation of the SLO process. The limited number of surveys that have been conducted show mixed and/or conflicting results about the type of feedback received and used by teachers from the SLO process (Diaz-Bilello & Thompson, 2019; Lachlan-Hache et al., 2015). In preparing this chapter, the authors found no studies on the effects of using SLOs on student achievement.

Evaluation of SLOs Informing Teaching and Learning

SLOs are a process designed to integrate assessment and instruction. The process involves teachers and administrators agreeing on specific learning goals that the teacher will monitor and on selecting assessments to use as measures over time of student progress on those goals. The process was largely overtaken by the teacher accountability movement. Little research to date has focused on the effectiveness of SLOs for improving teaching and learning.

Assessment Based on Learning Progressions

The National Research Council (2007, p. 219), focusing on science education, defined learning progressions as “descriptions of the successively more sophisticated ways of thinking about a topic that can follow one another as children learn.” Corcoran et al. (2009, p. 8) used a more expansive definition: “Learning progressions in science are empirically grounded and testable hypotheses about how students’ understanding of, and ability to use, core scientific concepts and explanations and related scientific practices grow and become more sophisticated over time, with appropriate instruction.” In mathematics, the term *trajectory* has also been used, stemming from Simon’s (1995, p. 135) proposal of a “hypothetical learning trajectory” that referred to mathematics teachers’ expectations for how student learning might proceed for certain mathematical content. Work in learning progressions has developed in other countries as well, beginning as early as the 1980s.

Although more work on learning progressions has been done in mathematics and science than in other disciplines, the idea has been applied to many different disciplines. The appeal is that learning progressions can form the basis for the teacher’s instructional decisions (Deane et al., 2015), improving teaching and learning. In fact, some definitions of learning progressions include not only progressions in students’ thinking but also progressions in instructional experiences that might help students move along, and the relative lack of attention to including the latter in progressions is one of the current criticisms of learning progression research (Mosher, 2022).

All definitions of learning progressions envision learning as growth in sophistication of understanding, not solely a body of content to be covered, and envision progress as growth in understanding not referenced to age or grade level (Heritage, 2008). Learning progressions describe typical development of expertise in a domain for most, but not

all, students. They are not rigid; a student's level on a learning progression may fluctuate as learning is taking place. In many domains, learning progressions may describe building blocks that lead up to more complete understandings. In science, lower levels on a particular learning progression may also describe commonly held facets or naive conceptions that students must abandon as their learning progresses (Alonzo, 2018). Since students learn in the context of instruction, learning progressions are not independent of curriculum sequence (Wiliam, 2018b).

Many—if not most—sets of content standards and curricula are built on a logical, developmental progression of knowledge, skills, and competencies across grade levels (e.g., Common Core Standards Writing Team, 2019; Confrey et al., 2014; National Centre, 2020) and certainly qualify as progressions in the plain English sense of the word. However, the empirically developed and validated learning progressions defined and described in this section represent something beyond this; while acknowledging a relationship with students' educational experiences, the assessments based on learning progressions described in this section also seek to detail cognitive growth in a more theoretical way.

Purpose and Uses of Assessment Based on Learning Progressions

Learning progressions are a hypothesized vehicle for making assessment more useful for informing teaching and learning, whether at the classroom or the large-scale level, and especially to guide teachers' formative assessment practices (Alonzo, 2018). Teachers need to have a conception of what comes next in learning to respond effectively to students' work. It is for this reason that informing teaching and learning seems the most salient purpose for learning progressions and assessments based on them. However, a learning progression framework can also be used to describe student performance summatively.

Design of Assessment Based on Learning Progressions

There are different approaches to constructing learning progressions as a priori, theoretical statements about how learning takes place in a domain. In general, such approaches are based on the work of human panelists (as opposed to data mining) and can be categorized as top-down, beginning with the ideas of experts in the domain, or bottom-up, beginning with the ideas of curriculum content experts and teachers about what is best taught when (Heritage, 2008). Learning progressions can be big, encompassing K–12 learning in a broad domain (e.g., writing), or relatively small, describing the acquisition of understanding on a single topic.

A true learning progression—as opposed to, for example, a curriculum scope and sequence or a simple rubric—requires empirical evidence or validation, or what Graf and van Rijn (2016) have called *empirical recovery*. If learning progressions are a representation of a theory about how children learn and develop in a domain, they should be empirically verifiable (Denvir & Brown, 1986; Graf & van Rijn, 2016).

Therefore, design concerns for assessment based on learning progressions include both verifying the progression and designing and validating measures to assess them. Confrey (2019) cautioned that studies need to distinguish between the learning progression and its measures. Evidence for the former must verify the constructs in the progression, while evidence for the latter must validate a scale for measuring student progress on the progression. When the two are conflated, what results is much like the validity evidence for any other educational measure. When the two are distinguished, research on the progression can inform research on the measure and vice versa.

Learning progressions design work is done in a content area, because learning progressions are content specific, and most often requires partnerships between researchers and educational practitioners. The result is often assessment that is neither completely classroom based and teacher developed nor completely externally developed, as for large-scale assessments. This hybrid development is a hallmark of assessments based on learning progressions, at least to date, and foreshadows some future possibilities as technology makes possible the coordination needed for this kind of assessment.

Measurement Considerations for Assessment Based on Learning Progressions

Some learning progressions are constructed to help teachers create or select classroom-based tasks and progressions, and some are constructed to serve as the basis for externally developed tasks and questions that teachers can use in the classroom. Content-related evidence for the validity of the progression has been reported for the former (Furtak, Circi, & Heredia, 2018; Hess & Kearns, 2010, 2011). A wider range of validity evidence has been reported for the latter, including content-related evidence as well as evidence of whether measurement models fit assessment data as predicted. It is worth noting that studies do not always support hypothesized learning progressions, or do not support them completely (e.g., Pham et al., 2021). Examples of how several different programs of research have treated validity and validation follow.

One of the more influential frameworks for the design of learning progressions (Duschl et al., 2011), especially in science, is the BEAR Assessment System (Berkeley Evaluation and Assessment Research Center; Wilson, 2009; Wilson & Sloane, 2000). The BEAR Assessment System is based on four principles, each with accompanying building blocks for the assessment system (Wilson et al., 2016, pp. 7–8):

- (1) Assessment should be based on a developmental perspective of student learning; the building block is a *construct map* of a progress variable that visualizes how students develop . . .
- (2) There must be a match between what is taught and what is assessed; the building block is the *items design* . . .
- (3) Teachers must be the managers of the system, with the tools to use it efficiently and effectively; the building block is the *outcome space*, or the set of categories of student responses that make sense to teachers . . .
- (4) There is evidence of quality in terms of reliability and validity studies and evidence of fairness, through the data analytics; the

building block is an algorithm specified as a *measurement model* that provides for a visual representation of the students and the items on the same graph (called a “Wright Map”) and a number of other data analytic tools.

The BEAR Assessment System, then, creates a priori learning progressions based on research and uses researcher-developed measures in a system managed but not authored by teachers (Black et al., 2011; Wilson & Scalise, 2016; Wilson & Sloane, 2000). The measures may be based on multiple-choice or open-ended items. Classroom teachers use the system during instruction. Validity can be supported if the instruments reflect the expected characteristics of learning described in the progressions. Construct maps can be used to study this match (Black et al., 2011), and statistics from unidimensional or multidimensional item response theory analyses can be offered as evidence for appropriate internal structure and reliability. Assessment results are used to adjust both the learning progression and the items (e.g., Schwartz et al., 2017). Table 15.3 shows an

Table 15.3 Construct Map Showing a Learning Progression for Student Understanding of Earth in the Solar System

Level	Description
5 8th grade	Student is able to put the motions of the Earth and Moon into a complete description of motion in the Solar System which explains: the day/night cycle the phases of the Moon (including the illumination of the Moon by the Sun) the seasons
4 5th grade	Student is able to coordinate apparent and actual motion of objects in the sky. Student knows that: the Earth is both orbiting the Sun and rotating on its axis the Earth orbits the Sun once a year the Earth rotates on its axis once per day, causing the day/night cycle and the appearance that the Sun moves across the sky the Moon orbits the Earth once every 28 days, producing the phases of the Moon Common error: Seasons are caused by the changing distance between the Earth and Sun. Common error: The phases of the Moon are caused by a shadow of the planets, the Sun, or the Earth falling on the Moon.
3	Student knows that: the Earth orbits the Sun the Moon orbits the Earth the Earth rotates on its axis However, student has not put this knowledge together with an understanding of apparent motion to form explanations and may not recognize that the Earth is both rotating and orbiting simultaneously. Common error: It gets dark at night because the Earth goes around the Sun once a day.

(continued)

Table 15.3 (continued)

Level	Description
2	Student recognizes that: the Sun appears to move across the sky every day the observable shape of the Moon changes every 28 days Student may believe that the Sun moves around the Earth. Common error: The Sun travels around the Earth. Common error: It gets dark at night because the Sun goes around the Earth once a day. Common error: The Earth is the center of the universe.
1	Student does not recognize the systematic nature of the appearance of objects in the sky. Students may not recognize that the Earth is spherical. Common error: It gets dark at night because something (e.g., clouds, the atmosphere, “darkness”) covers the Sun. Common error: The phases of the Moon are caused by clouds covering the Moon. Common error: The Sun goes below the Earth at night.
0	No evidence or off-track

Note. From “Measuring Progressions: Assessment Structures Underlying a Learning Progression,” by M. Wilson, 2009, *Journal of Research in Science Teaching*, 46(6), p. 720. Copyright 2009 by John Wiley and Sons. Reprinted with permission.

example of a construct map that functions as a learning progression for student understanding of the Earth in the solar system (Wilson, 2009, p. 720).

Confrey and her colleagues (Confrey & Toutkoushian, 2019; Confrey et al., 2020) created Math Mapper 6–8, a digital learning system used in middle-grades mathematics classrooms. Multiple forms of assessment are used, including multiple choice, multiple select, numeric, one letter, and open ended. Like the BEAR system, Math Mapper 6–8 is based on a priori learning progressions derived from research, but it uses a different validation framework. The validation process is iterative: Results from the digital learning system are used to adjust the learning trajectory, and vice versa. Evidence is sought that the assessment taps domain knowledge (cognitive evidence), that the assessment is aligned to the curriculum and opportunity to learn and is used to inform teaching and learning (instructional evidence), and that the assessment reliably yields model-based information about student performance (inferential evidence; Pellegrino et al., 2016). Both quantitative (e.g., from empirical studies of model fit) and qualitative (e.g., observation-based design studies) evidence are sought.

Pham et al. (2021) studied three learning progressions for formative assessment in middle school mathematics. They, too, used a priori learning progressions derived from research to describe three related progressions—equality and variable, linear functions, and proportional reasoning—and created items to measure them. They used item response theory to investigate whether student response

evidence supported claims about the order of levels within each progression and multidimensional item response theory to investigate whether student response evidence supported predicted relationships across pairs of progressions.

In language, Bailey and Heritage (2014) use a more descriptive and qualitative approach to validation of learning progressions in the Dynamic Language Learning Progression project. The project collected a body of elementary school-age children's oral and written language (including from English learners and children with a range of language experiences) and from this corpus created learning progressions illustrating how children's use of language in school becomes gradually more sophisticated over time. Validation included a study of students' oral explanations (both procedural and conceptual) of mathematics problems. The language samples are searchable online. The language learning progressions are intended for use with both regular and English learner students and eventually are intended to inform interim and summative assessment, as well as formative assessment and classroom instruction (Bailey & Heritage, 2014) and student self-assessment (Goral & Bailey, 2019).

One critique of the learning progressions approach is that cognitive understanding is not quite as linear as learning progressions may suggest (Shavelson, 2009). Several research agendas have used nonlinear cognitive models or maps that relate knowledge components within a content area. Automated methods for cognitive diagnostic modeling are often used to create these cognitive maps (Goldin et al., 2016). Wilson et al. (2016) saw a place for exploratory automated methods as well as confirmatory methods such as those used in the BEAR system, partly depending on how much is already known about learning in an area. However, some argue that even in well-researched areas such as mathematics learning, automated methods help with the diagnostic precision needed for teachers to make decisions about next instructional moves (Kingston & Broaddus, 2017; R. Liu & Koedinger, 2017).

For example, R. Liu and Koedinger (2017) showed that most students who can calculate the area of a two-dimensional shape given its sides can also calculate the sides, given the area, and human experts most often consider calculating area of a shape to be one skill. However, cognitive modeling showed that this is not true for all shapes: for example, calculating from the radius to the area of a circle is a different skill from calculating the radius given the area. R. Liu and Koedinger (2017) showed that is probably because of the use of the square root. This additional precision about student understanding has obvious utility for informing teaching and learning.

Another example of using automated techniques for cognitive mapping is the Dynamic Learning Maps project, which uses diagnostic classification modeling to create the learning maps (Kingston et al., 2016). Originally applied to alternate assessment of state standards, Dynamic Learning Maps can support formative assessment and instructional decisions about typical educational standards as well (Kingston & Broaddus, 2017) and allows for representing finer grained, more diagnostic information for content standards than that produced by large-scale, standards-based tests.

In sum, to date, much of the measurement consideration for assessment based on learning progressions has focused on the development and validation of the learning progressions themselves. Relatively little research has focused on traditional measurement considerations related to their use by teachers, schools, and districts for various formative or summative purposes. Learning progressions may eventually provide a framework for the design of assessment systems, for example, district assessment systems, that include both classroom and large-scale assessment (Duschl et al., 2011). The most effective assessment systems will relate formative and summative measures in coherent, comprehensive, and continuous ways (Wilson & Draney, 2004) that most current assessment systems do not.

Implementation of Assessment Based on Learning Progressions in Practice

Learning progressions and assessment based on them are still very much in the developmental or small-scale implementation phase. Such development and implementation almost always involves a research–practitioner partnership of some sort. Even so, teachers do not always use learning progressions as intended (Alonzo & Elby, 2019).

Australia has been an early adopter of this approach, beginning work on what they called “subject profiles” in 1988 (Rowe & Hill, 1996). Subject profiles are vertical maps of achievement in a domain that can be used as a framework for assessing student work. The work in Australia was conceived in a summative context, looking for ways to record and report student growth. Currently, it is also used formatively, illustrating the aforementioned trend toward assessment being called on to inform teaching and learning that is happening all over. Australia currently uses a learning progression approach to student growth in reform efforts aimed at both curriculum and assessment, for both formative and summative purposes (Commonwealth of Australia, 2018; Confrey, 2019).

Furtak and Heredia (2014; Heredia et al., 2016) showed that the design and use of common formative assessments using a learning progression can inform how teachers interpret student thinking (Furtak, Circi, & Heredia, 2018) and adjust instruction. Furtak, Circi, & Heredia (2018) showed that a group of science teachers were able to design formative assessment tasks generally aligned with multiple learning progressions in high school biology, specifically in the area of natural selection, with some gaps and variations and associated uneven student learning gains across the learning progressions. In this research agenda, implementation can also proceed through the external development of curriculum-embedded assessments of learning progressions (Briggs & Furtak, 2020), including performance-based tasks and labs, phenomenon-based item clusters, and conceptually oriented multiple-choice questions.

Math Mapper 6–8 (Confrey & Toutkoushian, 2019; Confrey et al., 2020) provides teachers with assessments designed to give immediate, actionable feedback to both

teachers and students. Teachers are allowed to use the assessments as they see fit during their instruction. Confrey and Toutkoushian (2019) found that teachers vary in their use of the assessments; most measure students' progress about two thirds of the way through an instructional unit, when there is still time to adapt instruction.

Black et al. (2011) described a broad model of pedagogy based on teachers' understanding "road maps" or progressions toward curricular learning goals and using classroom assessment strategies that provide feedback to students. Wilson and colleagues applied this model in the BEAR Assessment System by working with content area researchers in various disciplines. Work has been conducted in the areas of science (Wilson & Draney, 2004; Wilson & Sloane, 2000), scientific argumentation (Yao et al., 2015), structure of matter (Morell et al., 2017), data science (Arneson et al., 2019), and statistical reasoning (Lehrer et al., 2014; Wilson & Lehrer, 2021). The BEAR Assessment System embeds assessment in the classroom teaching and learning context. Teachers are involved in managing the way the assessment system is used in their classrooms, collecting and scoring student work, interpreting the results in instructional terms, and using the results in instructional decision-making. To facilitate this, the BEAR Assessment System provides a single scoring guide for each variable in the progression; teachers do not need to use a new scoring guide for each assessment. In so doing, the assessment gives direct evidence of students' location on the progression and implications for instruction. Looking beyond the development of assessments, Lehrer et al. (2014) developed an approach that integrates the learning progression for an instructional program in middle school statistics with predeveloped items and teacher-managed formative assessment. This approach has been expanded to incorporate a mobile device-based system for recording teachers' daily observations of their students in alignment with the postulated learning progression (Lehrer, 2021; Wilson, 2021, 2023).

Impact of Assessment Based on Learning Progressions on Learning

Most studies of the effectiveness of learning progressions have studied the learning progressions themselves, seeking to confirm that students do in fact demonstrate thinking as described in the levels, or studied their measures, seeking to validate that performance on a measure reflects the levels described in the learning progressions. Some have studied the effects of using learning progressions, or of professional development using learning progressions, on the alignment of teacher-created formative assessment tasks, teachers' interpretation of student work, or instructional decision-making (Alonzo & Elby, 2019; Furtak, Bakeman, & Buell, 2018; Heredia et al., 2016); in other words, these studies investigated how learning progressions might inform teaching.

Furtak, Circi, & Heredia(2018) showed that students whose biology teachers participated in professional development using learning progressions, designed formative

assessment, used it with their students, and reflected on next steps in instruction did increase along the learning progression as measured against their own baselines.

A few studies have used experimental designs. Wilson and Sloane (2000) used an experimental design to evaluate the initial work of the BEAR Assessment System in the context of middle school science. Results demonstrated that students whose teachers used the assessment system improved their learning—with both statistically and educationally significant effects—beyond that of students whose teachers were in either of two control groups: teachers who participated in professional development but were not required to use the assessment system and comparison teachers who taught the regular middle school science curriculum. More recently, BEAR Assessment System work in the context of middle school statistics also showed statistically and educationally significant results in a controlled experiment using a change score approach (Gochyyev & Wilson, 2021; Wilson & Lehrer, 2021).

Some experimental studies have investigated the impact of using cognitive tutoring systems based on cognitive maps and found the effect to be strong and positive (Ritter et al., 2007). R. Liu and Koedinger (2017) compared achievement of students who used a cognitive tutor with that of students who used the cognitive tutor redesigned with adjustments to the cognitive map based on the results of data mining. Students in the redesigned condition had significantly higher posttest scores than control students, even after controlling for pretest scores, which they interpreted to mean that the quality of the cognitive map had improved and made a difference in the effectiveness of the tutoring system.

Evaluation of Assessment Based on Learning Progressions Informing Teaching and Learning

The idea of learning progressions has both theoretical and intuitive, practical appeal. Logically, learning progressions should help address the known difficulty teachers have deciding precisely what feedback or next instructional moves would help learners the most in moving toward learning goals (Heritage et al., 2009; Ruiz-Primo & Furtak, 2006, 2007; Ruiz-Primo & Li, 2013; Schneider & Gowan, 2013; see the “Formative Assessment” section above). Teachers demonstrably do need some way to conceptualize “what’s next” to adjust their teaching.

The current state of the field of assessments based on learning progressions shows some promising proofs of concept. Most of these projects leverage the recent increase of technology in the classroom to support classroom-based formative assessment with deep pockets of external research and methodological capability. Several projects leverage teacher professional development and external research support, some with and some without large-scale measurement methodology. At present, it is difficult to imagine enough of these resource-intensive projects to cover the curriculum (also see Baird et al., 2017). When these projects talk about going to scale, they usually mean involving more classrooms in a computer-based tutoring or assessment system, rather than being generally or routinely used in the educational landscape.

Work on assessment based on learning progressions does illustrate the major theme of this chapter, the trend toward the formative. Learning progressions research continues the shift from the focus on groups and aggregate scores to the focus on individuals. Yet several of the research programs described in this section show how difficult it is to apply a particular learning progression to an individual student, even if it is clear what “typical development” looks like.

Work on learning progressions also highlights the role of learning theory in assessment and assessment design. Most of the learning progression research described in this section is theory driven, with the progression distilling research on how students learn particular content into a developmental path. This section shows that personalized learning and assessment system design fueled by data science also has the potential to be driven more by data than by theory. The balance of theory and data science in future learning progression development is likely to be quite interesting.

FUTURE DIRECTIONS

The concepts of teaching and learning continue to evolve, and the pace of change appears to be increasing (Bennett, 2014). In a historical sense, the inseparable connection between teaching and learning is still relatively new, with research on teaching that focused on the student dating only to the last quarter of the 20th century (Peterson, 1979; Rosenshine, 1979; Shulman, 1986) and the current focus on student-centered assessment emerging around the same time (Chappuis, 2022; Stiggins, 1994). The definition and scope of learning—and student-centered education in particular—continues to expand, redefining nearly all aspects of teaching and learning: what is taught, how it is taught, why it is taught, who is taught, and by whom (Carter & Darling-Hammond, 2016; Cochran-Smith & Lytle, 1999; Faltis & Valdés, 2016). As previously described, this shift from “instructivist” (Box et al., 2015, p. 972) or “social efficiency” (Shepard, 2000, p. 4) teaching focused on delivering content to learners to teaching that supports students as they construct their own meaning has fueled the press for all assessment to be formative. The argument will be summarized in more detail in the “Conclusion” section; here, it is enough to note the growth in interest and use of formative assessment methods (classroom formative, interim, and common formative), the use of summative assessment methods (grading and state accountability assessment) for formative purposes, and the development of assessments focused on trajectories that describe the process as well as the final outcome of learning.

In addition, the introduction of technology into almost every aspect of schooling, from record keeping to communication, to curricular and instructional resources, to instruction itself, is changing the nature of teaching and learning (Bennett, 2015, 2018; see also Bennett et al., this volume). Information technology is advancing rapidly, including the availability of computers and other devices in the classroom, as well as Internet and communications technology linking students in classrooms to external information sources and to other students. The rapid growth of information technology has generally

changed most aspects of schooling—for better or worse—and has the potential to both support and shape the ongoing transition to more learning-oriented assessment.

It is within this fluid environment that the future of assessment to inform teaching and learning will be determined. The argument throughout this chapter has been that usefulness of assessment in informing teaching and learning is directly related to its proximity to learning, factors such as the timing and nature of the content assessed and the manner in which it is assessed, the nature of the information the assessment provides as feedback to teachers and students, and the type and level of pedagogical skill needed to effectively use the information provided. It is inevitable that the nature of assessment information will change as the understanding of teaching and learning evolves (Mislevy, 2018a; Pellegrino, 2014; von Davier et al., 2019; Wilson & Scalise, 2016) and the constructs that are valued and measured shift to meet societal needs (L. Liu et al., 2017; Stoeffer et al., 2020). It is likely that prevailing measurement models and measurement principles will be adapted or replaced as the field expands its view of what is important to measure (Mislevy, 2018b; Taylor et al., 2018). And it is possible that the very nature of assessment will change as technological advances, including advances in artificial intelligence, increasingly reshape education and all aspects of life (Bennett, 2018; von Davier et al., 2021; Wilson et al., 2015).

The Future of Assessment Information

In the third edition of *Educational Measurement*, Bunderson et al. (1989) outlined a four-generation framework to describe the effect that ongoing advances in technology would have on educational assessment.

- *Generation 1. Computerized testing:* administering conventional tests by computer.
- *Generation 2. Computerized adaptive testing:* tailoring the difficulty or contents of the next item or an aspect of the timing of the next item on the basis of test takers' responses.
- *Generation 3. Continuous measurement:* using calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student's achievement trajectory and profile as a learner.
- *Generation 4. Intelligent measurement:* producing intelligent scoring, interpretation of individual profiles, and advice to learners and teachers, by means of knowledge bases and inferencing procedures (Bunderson et al., 1989, p. 368).

Further, Bunderson et al. characterized a “shift in emphasis from institutional purposes [of educational measurement] to individual purposes” as characterizing “the distinction between the first two and last two generations of educational measurement” (p. 369). It is such a shift in emphasis in the purpose of educational measurement that is likely to move the field from Generation 1 and 2 assessments, designed to provide information about groups of students to support policy decisions about institutions, to a much greater focus on Generation 3 and 4 assessments, designed to produce

information about individual student performance and growth to inform teaching and learning. For example, Zheng et al. (2019) showed that formative assessment information embedded in MATHia, an intelligent tutoring system for middle and high school blended learning, can predict state test scores beyond the ability of a prior-year test.

As demonstrated throughout this chapter, in the time between the third and fifth editions of *Educational Measurement*, the shift in demands on assessment from serving institutional purposes to serving individual purposes has begun and is exerting growing pressure on Generation 1 and 2 assessments and their stakeholders. While many classrooms currently use Generation 1 and 2 assessments, Generation 3 and 4 assessments are being gradually introduced in research projects such as the ones described in the section on “Assessment Based on Learning Progressions” and require partnerships between schools and external organizations. Generation 3 and 4 assessments are not yet the common practice—but then, not that long ago, neither were Generation 1 and 2 assessments.

Although the field as a whole has progressed through the framework over the 30 years since the framework was published, the past decade has seen significant movement in K–12 assessment. Advances in the capabilities and availability of technology have fueled widespread adoption of computer-based assessment and rapid acceleration from Generation 1 (computerized testing), through Generation 2 (computerized adaptive testing), to the cusp of Generation 3 (continuous measurement). Computerized adaptive testing is well established as the primary mode of testing for interim assessments and has been accepted as a viable option for state summative assessment (Bennett, 2015).

It is in the classroom, however, that there is most likely to be an emergence of Generation 3 assessment—the use of calibrated measures embedded in a curriculum to continuously and unobtrusively estimate dynamic changes in the student’s achievement trajectory and profile as a learner (Bennett & Gitomer, 2009; Scalise & Wilson, 2011; Wilson & Sloane, 2000). This brings with it implications for privacy rights and responsibilities, teacher assessment literacy, and school technological capacity and policies.

Constructs

The constructs measured by assessments designed to support teaching and learning are likely to shift in two areas. The first shift involves measuring content in currently measured academic areas to a much greater depth and measuring additional areas that are not currently a primary focus of K–12 educational measurement. The second shift involves measuring process as well as outcomes in each of those areas.

The first shift—driven by academic content, curricular, and instructional changes—involves the use of assessments designed to measure students’ higher order thinking skills via performance on complex tasks. The first shift also includes measurement of other skills not currently a primary focus of K–12 educational assessment. Those include

so-called 21st-century skills such as communication, teamwork, and collaboration, as well as constructs such as social-emotional learning and student engagement. While these types of changes in assessment are already being driven by the development of new content standards in the core areas of English language arts, mathematics, and science, it is likely that similar changes will also be seen in other content areas. These changes may bring with them long-standing generalizability issues related to performance assessment because responses to complex tasks are subject to a large person-by-task interactions (Gao & Baxter, 1994; McBee & Barnes, 1998).

The second shift involves increased interest in and capacity to measure process in addition to student outcomes. Computerized assessment has provided the opportunity to efficiently collect process data such as time spent completing an item or progress through an instructional unit, intermediate steps or pathways, and changes to responses, as well as student outcomes. Including such process data in measurement has the potential to improve significantly the quantity and quality of information about student thinking that can be derived from assessment. As discussed previously, for formative assessment the main goal is not measurement of student achievement of a learning goal, but rather a description of the status of student understanding on the way toward a learning goal; the main information sought is an understanding of the nature of student thinking, not the correct answers that thinking will generate (Heritage & Heritage, 2013).

The availability of additional data alone, however, is likely to do little to improve the ability or use of assessment to inform teaching and learning. It will be necessary to process that wealth of data to produce Generation 3 and 4 type information that helps teachers understand students' achievement trajectories and interpret individual student profiles or offers instructional advice to learners and teachers. Generating that type of information will require the expansion of current measurement models, as well as the development of new models, and will likely incorporate long-standing and emerging work and modeling in the areas outside traditional psychometrics and assessment, such as computer-assisted instruction, intelligent tutoring, and data mining (Gobert et al., 2013; Koedinger et al., 2018; Ritter et al., 2007).

Measurement and Other Models

Since the NCLB era began in 2001, the design and development of K–12 large-scale summative and interim assessments has relied on the application of unidimensional measurement models such as the Rasch model and a variety of one-, two-, and three-parameter item response theory models. The use of these models has served the field well for accomplishing the institutional purposes of educational measurement (i.e., providing comparable estimates of students' overall ability across a variety of alternate test forms administered within and across years). As discussed in the sections of this chapter on state summative and interim assessment, these models have been less useful in providing information to inform teaching and learning, particularly at the individual student level.

As the focus of educational measurement has shifted to the classroom and individuals, new models and different applications of current models have emerged to inform the design of assessments intended to inform teaching and learning and support formative assessment. (See Cai et al., this volume, for a discussion of modeling for different types of assessment.) Theory-driven applications of the traditional unidimensional and multidimensional Rasch models have been used to help define, describe, and measure student performance along learning progressions with the specific intent of providing information to inform teaching and learning (Black et al., 2011; Briggs & Alonzo, 2012; Shepard, 2018; Wilson, 2012). As discussed in the section on learning progressions, to date, much of the innovative research in this area is still being conducted in close partnership between researchers and local educators. Widespread application of this emerging work at scale has yet to be accomplished.

While the unidimensional and multidimensional applications of the Rasch model have been useful in describing relatively simple linear progressions, new models or modeling approaches are needed to support the development of assessments to measure and support the development of complex learning progressions/trajectories or learning maps (Confrey et al., 2014; Toutkoushian, 2019; Kingston & Broaddus, 2017; Kopriwa et al., 2016). In addition, latent class models such as cognitive diagnostic models and hierarchical diagnostic classification models have applied a multidimensional approach to describe more fine-grained profiles of student performance and estimates of mastery (Bradshaw et al., 2014; Templin & Bradshaw, 2014). In addition to these models focused on outcomes for individual students interacting with items and tasks, new measurement models are being developed to provide information for new types of assessment designed to measure 21st-century skills such as collaboration.

Computational psychometrics is described as

a blend of stochastic processes theory, computer science–based methods, and theory-based psychometric approaches that may aid the analyses of complex data from performance assessments, including collaborative assessments. This discipline organically developed around the complex next-generation learning and assessment systems that include performance tasks, such as collaboration, games, and simulations. (von Davier, 2017, p. 3)

Computational psychometrics combines theory-based psychometrics and computer science–based methods to develop a model that incorporates process and outcome data into information about student performance on complex tasks.

The computer science–based methods employed in computational psychometrics, such as machine learning and data mining, can also be applied on their own without theory-based psychometrics to develop models of student learning and achievement. Such models fall under the broader category of learning analytics, which may incorporate data mining, machine learning, artificial intelligence, and other data science techniques in an attempt to provide information to inform instruction. Although instructional

theory, learning theory, and measurement theory are not necessary aspects of learning analytics, it is expected that the most effective use of learning analytics to inform teaching and learning will involve an iterative process in which theory and empirical data are used to guide the collection and analysis of data and transform data into understanding that is useful to teachers and students (Wong et al., 2019). As Wilson and Scalise (2016, p. 2) noted, “applying learning analytics to educational assessment therefore requires negotiating a key intersection, which is at the interface of measurement technology and information technology.” At a minimum, it should be expected that the dissemination of information and feedback to teachers and students from learning analytic models and any of the models described above will be based on a solid theoretical foundation.

Nature of Assessment

As described throughout this chapter, the nature of formative assessment to inform teaching and learning is distinctly different from the nature of summative and interim assessment designed primarily to serve institutional purposes. By their nature, the process of formative assessment, and therefore the tools to support formative assessment, are most effective when they fit the Generation 3 description of measures that are embedded in a curriculum to estimate, continuously and unobtrusively, dynamic changes in the student’s profile as a learner. At this time, however, the field of educational measurement has not yet reached the point where technology can support continuous and unobtrusive assessment in the classroom. The role of the student in technologically based measurement also bears investigation. As advances in technology progress, it can be expected that some aspects of teacher and student activities and interactions in the classroom will be more likely candidates for continuous and unobtrusive data collection than others. As assessments and measurement models are built from that data, it will be critical to ensure that the data collected from the classroom are the most relevant data to inform teaching and learning and that they are representative of the interactions among teachers and students.

Assessment Literacy Requirements in the Future

Teacher assessment literacy requirements for the effective use of assessment information will remain centered on deep content knowledge and the attendant understanding of how to select learning targets and success criteria, communicate them to students, and accurately recognize how students are thinking from the evidence in their work. Teachers will benefit from deeper content and pedagogical knowledge as content, curriculum, and instructional changes focus teaching and learning on more complex cognitive processes and greater individualization and personalization.

An additional assessment literacy demand is likely, namely, that teachers will have the skills needed to filter more data and information than is currently normally available for individual students. At this time, teachers are often asked to make sound instructional decisions based on sparse evidence. As information technology makes

the collection, processing, and storage of data from the classroom feasible, teachers will have to evaluate multiple sources of information and data, identify relevant information, and combine multiple pieces of data to make informed decisions. They will need supports to do this (U.S. DOE, 2009). An alternative view, from the artificial intelligence research, is that the teacher's job becomes simpler, with the technology doing the interpretation and presenting the teacher with a simple set of options (Hamilton et al., 2023). The authors doubt that this will be the case. As teaching has become more and more complex, the central role of the teacher as the most influential factor in students' learning has been demonstrated over and over (Hanushek, 2011; Ker, 2016; Oppen, 2019).

As the field transitions from Generation 3 to Generation 4 measurement and the information provided from assessment becomes more prescriptive, some assessment literacy demands placed on teachers may be reduced; however, demands for skills in other areas are likely to increase. Teachers will need to be able to judge the quality of the information provided and determine the basis of the information: Was it customized for an individual student or generic for a wide variety of students? Was it based on learning theory or solely on probability?

Student assessment literacy requirements will remain focused on the ability to understand learning targets and criteria, compare work to criteria, and use different types of feedback to improve. As instruction becomes more personalized through the use of technology, students are likely to be required to process feedback from sources other than their teacher. Similarly, an increased focus on collaborative skills may require students to process real-time feedback from their peers in the course of completing performance-based activities.

Advances in ways researchers and educators can work together to determine what data to collect, what information to convey to teachers and students, and the best ways to convey that information will be critical to advancing the assessment literacy of teachers and students and improving the usefulness of information from assessment to inform teaching and learning.

CONCLUSION

The chapter has explored eight general types of assessment—classroom formative assessment, interim assessment, common formative assessments, grading, state summative assessment, curriculum-based measurement, student learning objectives, and assessments based on learning progressions—each used in different ways to inform teaching and learning. These types of assessment vary in the nature of the assessment information they provide, their effectiveness in terms of improvements in teaching and learning, the quality of the assessment process and information, and assessment literacy requirements. Expectations about the future of assessment to inform teaching and learning suggest that these types are changing and will continue to change.

All told, descriptions of these eight types of assessment work together to demonstrate the movement that Bunderson et al. (1989) anticipated and recent history has borne out. The arc of the history of assessment is moving from institutional purposes and bending toward assessment to inform teaching and learning.

At this time, the argument is not that the formative trend in assessment is largely successful in informing teaching and learning. The chapter has documented both progress and significant challenges. Rather, the argument is that assessments are being pressed to do so. Four of the types of assessment discussed in this chapter—classroom formative assessment, curriculum-based assessment, common formative assessments, and assessments based on learning progressions—were developed specifically to inform teaching and learning. Interest in classroom formative assessment especially is growing, and research suggests that, when done well, it can be highly effective. The other four types—student learning objectives, interim assessment, grading, and state summative assessment—were developed for more institutional purposes and bent toward the purpose of informing teaching and learning as a response to stakeholder and public demand. As the chapter has shown, that bending produced some cracks in validity and reliability, and efficacy research, where available, has shown that the latter four types of assessment are less effective for producing learning gains than are the types of assessment designed for that purpose.

Nevertheless, both educators and the public now believe that wherever information about student achievement is available, it should be harnessed in the service of teaching and learning, and they press for that at every turn. This chapter has discussed some of that pressure, including pressure on tests at all levels to produce formative information, the relatively recent development of adding a layer of district-mandated assessment (e.g., interim and common formative assessment) in order to have more assessment information, and the rise of a grading reform movement that for the first time explicitly states that grading should support learning. The conviction that assessment should inform teaching and learning is expressed on state and district websites and materials, in meetings at all levels of educational administration, in the classroom, and in communication with parents and caregivers.

Related to this movement toward the formative, assessment to inform teaching and learning is moving toward more student involvement. Pressure toward more student involvement comes from the advent of more situated, sociocognitive, and sociocultural views of learning, as well as concerns for students' social-emotional well-being, social justice, and fairness. As it has become clearer that students construct knowledge with the help of teachers, peers, and materials, and that students regulate their learning, it follows that students need to be involved in the generation and interpretation of evidence of that learning. Research has affirmed this general trend to be useful for informing teaching and learning. For example, feedback research has shown that elaborated feedback on which students can build is generally more effective than simple knowledge of results. Formative assessment research also has shown that students do better

when they understand the criteria by which they will be evaluated and use them as learning is taking place.

Therefore, informing teaching and learning means designing, administering, and using information that is, to the extent possible, student specific, producing valid and reliable information about individual learners as well as groups; pedagogy and content specific, producing information at instructionally actionable grain sizes; and situation specific, producing information relevant to a learning community, most often a classroom. This is not to say that assessment should not continue to serve institutional, accountability, or administrative purposes or produce valid and reliable data for groups relevant to administrative decision-making—for example, schools, districts, or states and demographic groups within those. The research reviewed in this chapter suggests that those purposes are being—and should be—overtaken, not eliminated, by purposes about supporting teaching and learning. This is happening in both effective and ineffective ways, as the reviews above described. The field of educational measurement needs to understand this phenomenon and continue to work to support quality assessment in a world where the balance of assessment information tips toward assessment to inform teaching and learning.

Points from this review may suggest ways to do this. However, these changes are not guaranteed, or even likely, to occur without concerted and coordinated effort among measurement professionals, content specialists, learning theorists, practicing educators, and others. As this review showed, progress is being made, but it is not easy, quick, or without challenges. With this caution, this review suggests three potential ways forward toward the intelligent assessment Bunderson et al. (1989) and others aim for.

First, the nature of assessment information needs to be understood in broader terms than conventional scales. At least one type of assessment, classroom formative assessment, includes verbal information, in the form of feedback comments, that is important for teaching and learning. In addition, verbal descriptions add meaning to performance levels in rubrics and developmental levels in learning progressions. Learning analytics can add process information, and such multiple measures can contribute to more precise estimates of student learning. This attention to meaning is not a new concept in educational measurement. The importance of inferences, interpretation, and the utility of tests and test scores has been a central tenet of educational measurement for decades. However, the movement toward the formative increases the importance of the underlying meaning of assessment information and the necessity of connecting interpretations of assessment results to understandings of content and pedagogy.

Second, attention to both the quality of the assessment process and the quality of the information will help with understanding and improving assessment to inform teaching and learning. Formative assessment, on the one hand, typically emphasizes the process, and more attention to the quality of the questions or discourse used in the formative assessment process might improve its effectiveness. Interim assessment, on the other

hand, typically emphasizes the quality of the information and has not yet really found an effective process for using that information in the service of teaching and learning. More work on this point may lead to diminished interest in interim assessments or more effective processes or designs for their use in teaching and learning. Currently, assessments or assessment systems that show promise for continuous or intelligent assessment are demonstration projects or successful research projects that explore assessment processes and tools that are far from commonplace in schools. In any case, insistence on quality of both process and information for all assessment information can be expected to improve teaching and learning.

Third, the more assessment turns away from institutional purposes and moves toward assessment to inform teaching and learning, the more assessment literacy demands rise for educators and students. In large-scale testing for institutional purposes, educators and students are consumers, and the bulk of assessment literacy requirements reside with test developers. As the trend moves toward assessment to inform teaching and learning, assessment users—educators and students—have a larger role to play and thus greater assessment literacy needs. This is the case even with professionally developed assessments that are used to inform teaching and learning, because the teachers and learners become the owners of the information and its use and consequences. It is even more the case with educator-developed assessments that are used to inform teaching and learning, for example, classroom formative assessment, common formative assessments, and grading. As the connection between assessment and the teaching and learning process expands, assessment literacy expands to include deep content knowledge and an understanding of how learning happens, in general and for specific concepts.

As assessment to inform teaching and learning evolves in these ways, as the chapter sections have suggested, advances will be required in validity theory and methods, measurement models and methods, technology, and assessment literacy. This is a broad list offering a wide array of productive opportunities for scholarship and practice and is not offered lightly—as this chapter shows, bumps in the road may be expected and collaborative work will be needed.

Finally, it is important to note that this chapter about assessment to inform teaching and learning has been written from an assessment point of view. The reviews have described scholarship and practice in eight types of assessment, described a broad assessment trend toward the formative, and derived three assessment-related sub-themes. The authors envision a companion chapter that could have been written by experts in teaching and learning, from a teaching and learning point of view. In other words, given the trend toward the formative, assessment to inform teaching and learning will also need to be understood from a learning point of view. Some of the research agendas reviewed have already resulted in partnerships with instructional experts and learning theorists. Advances in this kind of work may truly be the future of assessment to inform teaching and learning.

ACKNOWLEDGMENTS

The authors express sincere gratitude to Drew Gitomer, Dylan Wiliam, and Mark Wilson, whose thoughtful reviews helped us immensely. The reviewers were instrumental, especially in organizing and presenting the vast amount of material covered in this chapter and in contributing suggestions from their deep subject matter expertise. We would also like to thank the editors of this volume, Linda Cook and Mary Pitoniak, for their wise leadership and guidance throughout this enormous undertaking.

REFERENCES

- Abrams, L. M., & McMillan, J. H. (2013). The instructional influence of interim assessments: Voices from the field. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment* (pp. 105–133). Information Age Publishing.
- Abrams, L. M., McMillan, J. H., & Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Educational Assessment, Evaluation and Accountability*, 27, 347–375. <https://doi.org/10.1007/s11092-015-9214-9>
- Abrams, L., Varier, D., & Jackson, L. (2016). Unpacking instructional alignment: The influence of teachers' use of assessment data on instruction. *Perspectives in Education*, 34(4), 15–28. <http://dx.doi.org/10.18820/2519593X/pie.v34i4.2>
- Achieve, Inc. (2013). *Proficient vs. prepared: Disparities between state tests and the 2013 National Assessment of Educational Progress (NAEP)*. <https://issuu.com/achieveinc/docs/naepbrieffinal05141594b2daa338c234/8>
- Ainsworth, L., & Viegut, D. (2006). *Common formative assessments: How to connect standards-based instruction and assessment*. Corwin.
- Allal, L. (2010). Assessment and the regulation of learning. In P. Peterson, E. Baker, & B. McGaw, (Eds.), *International Encyclopedia of Education* (Vol. 3), pp. 348–352. Elsevier.
- Allal, L. (2011). Pedagogy, didactics and the co-regulation of learning: A perspective from the French-language world of educational research. *Research Papers in Education*, 26(3), 329–336. <http://dx.doi.org/10.1080/02671522.2011.595542>
- Allen, A., & Smith, R. A. (2022). Curriculum-based measurement. In D. Fisher (Ed.), *Routledge resources online: Education*. <https://doi.org/10.4324/9781138609877-REE114-1>
- Alonzo, A. C. (2018). An argument for formative assessment with science learning progressions. *Applied Measurement in Education*, 31(2), 104–112. <https://doi.org/10.1080/08957347.2017.1408630>
- Alonzo, A. C., & Elby, A. (2019). Beyond empirical adequacy: Learning progressions as models and their value for teachers. *Cognition and Instruction*, 37(1), 1–37. <https://doi.org/10.1080/07370008.2018.1539735>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84(3), 261–271. <https://doi.org/10.1037/0022-0663.84.3.261>
- Anderson, L. W. (2018, April). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives*, 26(49). <https://epaa.asu.edu/index.php/epaa/article/view/3814/2053>
- Andrade, H. L. (2013). Classroom assessment in the context of learning theory and research. In J. H. McMillan (Ed.), *Handbook of research on classroom assessment* (pp. 17–34). Sage.
- Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment in Education*, 27(4), 350–372. <https://doi.org/10.1080/0969594X.2019.1571992>
- Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education*, 17(2), 199–214. <http://dx.doi.org/10.1080/09695941003696172>
- Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary students' writing. *Educational Measurement: Issues and Practice*, 27(2), 3–13. <https://doi.org/10.1111/j.1745-3992.2008.00118.x>
- Andrade, H. L., Hefferen, J., & Palma, M. (2019). Formative assessment in the arts. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 126–145). Routledge Taylor & Francis.
- Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice*, 48(1), 12–19. <https://doi.org/10.1080/00405840802577544>
- Arneson, A., Wihardini, D., & Wilson, M. (2019). Assessing college-ready data-based reasoning. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 93–120). Routledge.
- Assessment Reform Group. (2002). *Assessment is for learning: 10 principles*. <https://www.aaia.org.uk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf>
- Au, W. (2007). High-stakes testing and curricular control: A qualitative meta-synthesis. *Educational Researcher*, 36(5), 258–267. <https://doi.org/10.3102/0013189X07306523>
- Austin Independent School District. (2015). *Student learning objectives (SLOs): Analysis of student growth in 2013–2014, by type and source of assessment*. Department of Research and Evaluation, Austin Independent School District. https://www.austinisd.org/sites/default/files/dre-surveys/DRE_14.85_Analysis_of_Student_Growth_in_2013_2014_by_Type_and_Source_of_Assessment.pdf
- Babo, G., Tienken, C. H., & Gencarelli, M. A. (2014). Interim testing, socio-economic status, and the odds of passing Grade 8 state tests in New Jersey. *RMLE Online*, 38(3), 1–9. <https://doi.org/10.1080/19404476.2014.11462116>

- Bailey, A. L., & Heritage, M. (2014). The role of language learning progressions in improved instruction and assessment of English language learners. *TESOL Quarterly*, 48(3), 480–506. <https://doi.org/10.1002/tesq.176>
- Baird, J.-A., Andrich, D., Hopfenbeck, T., & Stobart, G. (2017). Metrology of education. *Assessment in Education*, 24(3), 463–470. <https://doi.org/10.1080/0969594X.2017.1337628>
- Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407. <https://doi.org/10.1177/0022487108324554>
- Balloo, K., Evans, C., Hughes, A., Zhu, X., & Winstone, N. (2018). Transparency isn't spoon-feeding: How a transformative approach to the use of explicit assessment criteria can support student self-regulation. *Frontiers in Education*, 3(69). <https://doi.org/10.3389/feduc.2018.00069>
- Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Educational Assessment, Evaluation and Accountability*, 22, 53–72. <https://doi.org/10.1007/s11092-010-9091-1>
- Banker, H. J. (1927). The significance of teachers' marks. *The Journal of Educational Research*, 16(3), 159–171. <https://doi.org/10.1080/00220671.1927.10879778>
- Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, 34(4), 31–40. <https://doi.org/10.1111/emip.12096>
- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education*, 18(1), 5–25. <https://doi.org/10.1080/0969594X.2010.513678>
- Bennett, R. E. (2014). Preparing for the future: What educational assessment must do. *Teachers College Record*, 116(11). <https://www.learntechlib.org/p/155766/>
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370–407. <https://doi.org/10.3102/0091732X14554179>
- Bennett, R. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice*, 37(4), 7–15. <https://doi.org/10.1111/emip.12231>
- Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Open University Press.
- Black, P., McCormick, R., James, M., & Pedder, D. (2006). Learning how to learn and assessment for learning: A theoretical inquiry. *Research Papers in Education*, 21(2), 119–132. <https://doi.org/10.1080/02671520600615612>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>

- Black, P., Wilson, M., & Yao, S. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, 9, 71–123. <https://doi.org/10.1080/15366367.2011.591654>
- Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education*, 85(2), 205–225. <https://doi.org/10.1080/01619561003685379>
- Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw–Hill.
- Boekaerts, M., de Koning, E., & Vedder, P. (2006). Goal-directed behavior and contextual factors in the classroom: An innovative approach to the study of multiple goals. *Educational Psychologist*, 41(1), 33–51. https://doi.org/10.1207/s15326985ep4101_5
- Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration*, 47(5), 609–629. <https://doi.org/10.1108/09578230910981080>
- Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation*, 17(3), 141–159. <https://dx.doi.org/10.1080.13803611.2011.597112>
- Box, C., Skoog, G., & Dabbs, J. M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal*, 52(5), 956–983. <https://doi.org/10.3102/0002831215587754>
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understanding of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33(1), 2–14. <https://doi.org/10.1111/emip.12020>
- Braun, H., Chapman, L., & Vezzu, S. (2010). The Black–White achievement gap revisited. *Education Policy Analysis Archives*, 18(21). <http://epaa.asu.edu/ojs/article/view/772>
- Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 293–316). Sense Publishers.
- Briggs, D. C., Chattergoon, R., & Burkhardt, A. (2019). Examining the dual purpose use of student learning objectives for classroom assessment and teacher evaluation. *Journal of Educational Measurement*, 56(4), 686–714. <https://doi.org/10.1111/jedm.12233>
- Briggs, D. C., & Furtak, E. M. (2020). Learning progressions and embedded assessment. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 146–169). Routledge.
- Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education*, 7(4), 279–301. https://doi.org/10.1207/s15324818ame0704_2

- Brookhart, S. M. (2001). Successful students' formative and summative use of assessment information. *Assessment in Education*, 8(2), 153–169. <https://doi.org/10.1080/09695940123775>
- Brookhart, S. M. (2013). Grading. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 257–271). Sage.
- Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment*, 20(4), 268–296. <http://dx.doi.org/10.1080/10627197.2015.1093928>
- Brookhart, S. M. (2017). Formative assessment in teacher education. In D. J. Clandinin & J. Husu (Eds.), *International handbook of research on teacher education* (pp. 927–943). Sage.
- Brookhart, S. M. (2018). Summative and formative feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 52–78). Cambridge University Press.
- Brookhart, S. M. (2020). Feedback and measurement. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 63–78). Routledge.
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86(4), 803–848. <https://doi.org/10.3102/0034654316672069>
- Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (REL 2007–No. 017). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2007017_sum.pdf
- Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.
- Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education*, 85(2), 186–204. <https://doi.org/10.1080/01619561003685346>
- Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 367–407). Macmillan.
- Burch, P. E. (2006). The new educational privatization: Educational contracting and high-stakes accountability. *Teachers College Record*, 108(12), 2582–2610. <https://doi.org/10.1111/j.1467-9620.2006.00797.x>
- Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education*, 85(2), 147–162. <https://doi.org/10.1080/01619561003685288>
- Burkhardt, H., & Schoenfeld, A. (2019). Formative assessment in mathematics. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 35–67). Routledge.

- Butler, D. L., & Schnellert, L. (2015). Success for students with learning disabilities: What does self-regulation have to do with it? In T. Cleary (Ed.), *Self-regulated learning interventions with at-risk youth: Enhancing adaptability, performance, and well-being* (pp. 89–112). APA Press.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65(3), 245–281. <https://doi.org/10.3102/00346543065003245>
- Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 71–84). Sage.
- Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing*, 26(3), 431–452. <https://doi.org/10.1007/s11145-012-9375-6>
- Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International*, 44(1), 57–66. <https://doi.org/10.1080/14703290601081332>
- Carless, D., & Boud, D. (2018) The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378–398. <https://doi.org/10.3102/0162373711412765>
- Carter, P. L., & Darling-Hammond, L. (2016). Teaching diverse learners. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 593–638). American Educational Research Association.
- Chapman, L. H. (2014). *The marketing of SLOs: 1999–2014*. <http://vamboozled.com/laura-chapman-slos-continued/>
- Chappuis, J. (2022). Student involvement in assessment. In D. Fisher (Ed.), *Routledge resources online: Education*. <https://doi.org/10.4324/9781138609877-REE8-1>
- Chojnacki, G., Eno, J., Liu, F., Meyers, C., Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013, September 26–28). *Do interim assessments influence instructional practice in year one? Evidence from Indiana's elementary school teachers* [Paper presentation]. Society for Research on Educational Effectiveness Annual Meeting, Washington, DC, United States. <https://files.eric.ed.gov/fulltext/ED563059.pdf>
- Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). *Making the most of interim assessment data: Lessons from Philadelphia*. Research for Action. <https://eric.ed.gov/?id=ED505863>
- Cizek, G. J., Andrade, H. L., & Bennett, R. E. (2019). Formative assessment: History, definition, and progress. In H. L. Andrade, R. E. Bennett, & G. J. Cizek (Eds.), *Handbook of formative assessment in the disciplines* (pp. 3–19). Routledge.
- Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence Public Schools* (WCER Working Paper No. 2008–10). Wisconsin Center

- for Education Research. https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2008_10.pdf
- Cochran-Smith, M., & Lytle, S. L. (1999). Relationships of knowledge and practice: Teacher learning in communities. *Review of Research in Education*, 24(1), 249–305. <https://doi.org/10.3102/0091732X024001249>
- Coe, M., Hanita, M. Nishioka, V., & Smiley, R. (2011, December). *An investigation of the impact of the 6+1 Trait Writing Model on grade 5 student writing achievement*. Institute of Education Sciences Report NCEE 2012-4010. https://ies.ed.gov/ncee/edlabs/projects/rct_52.asp?section=ALL
- Common Core Standards Writing Team. (2019). *Progressions for the Common Core State Standards for Mathematics (draft February 7, 2019)*. Institute for Mathematics and Education, University of Arizona.
- Commonwealth of Australia. (2018, March). *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools*. https://docs.education.gov.au/system/files/doc/other/662684_tgta_accessible_final_0.pdf
- Community Training and Assistance Center. (2004). *Catalyst for change: Pay for performance in Denver. Final report*. <https://ctacusa.com/wp-content/uploads/2013/11/CatalystForChange.pdf>
- Community Training and Assistance Center. (2013). *It's more than money: Teacher incentive fund—leadership for educators' advanced performance Charlotte–Mecklenburg Schools*. <https://ctacusa.com/wp-content/uploads/2013/11/MoreThanMoney.pdf>
- Community Training and Assistance Center. (2018). *Student learning objectives (SLOs)*. <http://www.ctacusa.com/education/student-learning-objectives-slos/>
- Confrey, J. (2019, May). *Future of education and skills 2030: Curriculum analysis: A synthesis of research on learning trajectories/progressions in mathematics*. Organisation for Economic Co-operation and Development.
- Confrey, J., Maloney, A., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM—The International Journal on Mathematics Education*, 46(5), 719–733. <https://doi.org/10.1007/s11858-014-0598-7>
- Confrey, J., & Toutkoushian, E. (2019). A validation approach to middle-grades learning trajectories within a digital learning system applied to the “Measurement of Characteristics of Circles.” In J. Bostic, E. Krupa, & J. Shih (Eds.), *Quantitative measures of mathematical knowledge: Researching instruments and perspectives* (pp. 67–92). Routledge.
- Confrey, J., Toutkoushian, E., & Shah, M. (2020). Working at scale to conduct ongoing validation of learning trajectory-based classroom assessments for middle grade mathematics. *Journal of Mathematical Behavior*, 60. <https://doi.org/10.1016/j.jmathb.2020.100818>
- Corcoran, T., Mosher, F. A., & Rogat, A. (2009, May). *Learning progressions in science: An evidence-based approach to reform* (CPRE Research Report No. RR-63). Consortium for Policy Research in Education. <https://files.eric.ed.gov/fulltext/ED506730.pdf>

- Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement* (NCEE 2013–4000). U.S. Department of Education. https://ies.ed.gov/ncee/rel/regions/midwest/pdf/REL_20134000.pdf
- Cotton, K. (1988, May). *Close-up #5: Classroom questioning*. Education Northwest. <https://educationnorthwest.org/sites/default/files/classroom-questioning.pdf>
- Council of Chief State School Officers. (2013). *Criteria for procuring and evaluating high-quality assessments*. <https://ccsso.org/resource-library/criteria-procuring-and-evaluating-high-quality-assessments>
- Council of Chief State School Officers. (2018). *Revising the definition of formative assessment*. <https://ccsso.org/sites/default/files/2018-06/Revising%20the%20Definition%20of%20Formative%20Assessment.pdf>
- Covington, M. V. (1992). *Making the grade: A self-worth perspective on school reform*. Cambridge University Press.
- Crooks, A. D. (1933). Marks and marking systems: A digest. *Journal of Educational Research*, 27(4), 259–272. <https://doi.org/10.1080/00220671.1933.10880402>
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438–481. <https://doi.org/10.3102/00346543058004438>
- Crouse, K., Gitomer, D. H., & Joyce, J. (2016). An analysis of the meaning and use of student learning objectives. In K. K. Hewitt, & A. A. Beardsley (Eds.), *Student growth measures in policy and practice* (pp. 203–222). Springer.
- Cushing, E., & Meyer, C. (2014). *Balancing autonomy and comparability: State approaches to assessment selection for student learning objectives*. *Ask the team*. Center on Great Teachers and Leaders, American Institutes for Research. <https://files.eric.ed.gov/fulltext/ED553372.pdf>
- Dadey, N., & Diggs, C. R. (2019, September). *A rapid review of interim assessment use*. National Center for the Improvement of Educational Assessment. <https://www.nciea.org/library/a-rapid-review-of-interim-assessment-use/>
- Dadey, N., Evans, C. M., & Lorie, W. (2023). Through-year assessment: Ten key considerations. The National Center for the Improvement of Educational Assessment. <https://www.nciea.org/wp-content/uploads/2023/03/Ten-Key-Considerations-Through-Year-Assessment-Report-March2023-F.pdf>
- Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. Teachers College Press.
- Davidson, K. L., & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. (CRESST Report 806). University of California, National Center for Research on Evaluation, Standards, and Student Testing. <https://cresst.org/wp-content/uploads/R806.pdf>
- Davis, B. (1997). Listening for differences: An evolving conception of mathematics teaching. *Journal for Research in Mathematics Education*, 28(3), 355–376. <https://doi.org/10.5951/jresmetheduc.28.3.0355>

- Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015, December). *Key practices in the English Language Arts (ELA): Linking learning theory, assessment, and instruction* (ETS Research Report No. RR-15-17). ETS. <http://dx.doi.org/10.1002/ets2.12063>
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35(1), 13–21. <https://doi.org/10.1016/j.intell.2006.02.001>
- Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, 52(6), 1133–1159. <https://doi.org/10.3102/0002831215596412>
- DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education*, 34(5–6), 576–591. <https://doi.org/10.1080/01626620.2012.730347>
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52(3), 219–232. <https://doi.org/10.1177/001440298505200303>
- Deno, S. L. (1993). Curriculum-based measurement. In J. C. Conoley & J. J. Kramer (Eds.), *Curriculum-based measurement—complete work* (pp. 1–23). Buros Institute of Mental Measurements.
- Deno, S. L. (2003a). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, 28(3–4), 3–12. <https://doi.org/10.1177/073724770302800302>
- Deno, S. L. (2003b). Developments in curriculum-based measurement. *The Journal of Special Education*, 37(3), 184–192. <https://doi.org/10.1177/00224669030370030801>
- Deno, S. L. (2013). Problem-solving assessment. In R. Brown-Chidsey & K. J. Andren (Eds.), *Assessment for intervention: A problem-solving approach* (2nd ed., pp. 10–36). Guilford Press.
- Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Minneapolis Leadership Training Institute. University of Minnesota. <https://files.eric.ed.gov/fulltext/ED144270.pdf>
- Denver, B., & Brown, M. L. (1986). Understanding of number concepts in low-attaining 7–9 year olds: Part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics*, 17(1), 15–36. <https://doi.org/10.1007/BF00302376>
- Diaz-Bilello, E., Burkhardt, A., & Briggs, D. (2016). Findings from case studies of the student learning objectives implementation experiences at four schools. Brief commissioned by the Denver Public Schools. Boulder, CO: Center for Assessment Design Research and Evaluation (CADRE). <https://www.colorado.edu/cadre/sites/default/files/attached-files/cadre-brief2.pdf>
- Diaz-Bilello, E., & Thompson, J. (2019). *Student learning objectives: Where are we now? Synthesis of studies examining the impact of student learning objectives on teaching*

- practices* [Unpublished manuscript]. Center for Assessment Design Research and Evaluation (CADRE).
- DuFour, R., DuFour, R., Eaker, R., & Karhanek, G. (2010). *Raising the bar and closing the gap*. Solution Tree Press.
- Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, 4(1), 37–73. https://doi.org/10.1207/s15326977ea0401_2
- Duschl, R., Maeng, S. & Sezen, A. (2011) Learning progressions and teaching sequences: a review and analysis, *Studies in Science Education*, 47(2), 123–182. <https://doi.org/10.1080/03057267.2011.604476>
- English, D., Ennis, J., Chamber, D., & Lachlan-Hache, L. (2015). *TIF 4 student learning objective review: Results and recommendations*. Maine Schools for Excellence. http://www.sad44.org/pdf/2015_Maine_SLO_Review_Report.pdf
- Espin, C. L., & Wallace, T. (2004). *Descriptive analysis of curriculum-based measurement literature* [Working document]. University of Minnesota Institute for Research on Progress Monitoring.
- Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & de Rooij, M. (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice*, 32(1), 8–21. <https://doi.org/10.1111/ldrp.12123>
- Espin, C. A., Saab, N., Pat-El, R., Boender, P. D., & van der Veen, J. (2018). Curriculum-based measurement progress data: Effects of graph pattern on ease of interpretation. *Zeitschrift für Erziehungswissenschaft*, 21(4), 767–792. <https://doi.org/10.1007/s11618-018-0836-9>
- Espin, C. A., van den Bosch, R. M., van der Liende, M., Rippe, R. C., Beutick, M., Langa, A., & Mol, S. E. (2021). A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision-making? *Journal of Learning Disabilities*, 54(4), 256–268. <https://doi.org/10.1177/0022219421997103>
- Espin, C., Chung, S., Foegen, A., & Campbell, H. (2018). Curriculum-based measurement for secondary-school students. In P. C. Pullen & M. J. Kennedy (Eds.), *Handbook of response to intervention and multi-tiered systems of support* (pp. 291–315). Routledge.
- Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat.1177 (2015–2016).
- Faltis, C. J., & Valdés, G. (2016). Preparing teachers for teaching in and advocating for linguistically diverse classrooms: A *vade mecum* for teacher educators. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 549–592). American Educational Research Association.
- Faria, A., Heppen, J., Li, Y., Stachel, S., Jones, W., Sawyer, K., Thomsen, K., Kutner, D., Miser, D., Lewis, S., Casserly, M., Simon, C., Uzzell, R., Corcoran, A., & Palacios, M. (2012, Summer). *Charting success: Data use and student*

- achievement in urban schools*. Council of the Great City Schools. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Charting_Success.pdf
- Farrell, C. C., & Marsh, J. A. (2016). Metrics matter: How properties and perceptions of data shape teachers' instructional responses. *Educational Administration Quarterly*, 52(3), 423–462. <https://doi.org/10.1177/0013161X16638429>
- Faxon-Mills, S., Hamilton, L. S., Rudnick, M., & Stecher, B. M. (2013). *New assessments, better instruction? Designing assessment systems to promote instructional improvement*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR354.html
- Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment and Evaluation in Higher Education*, 36(1), 51–62. <https://doi.org/10.1080/02602930903197883>
- Ferrara, S., Maxey-Moore, K., & Brookhart, S. M. (2019). Guidance in the *Standards* for classroom assessment: Useful or irrelevant? In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and measurement* (pp. 97–119). Routledge.
- Foegen, A. (2008). Algebra progress monitoring and interventions for students with learning disabilities. *Learning Disability Quarterly*, 31(2), 65–78. <https://doi.org/10.2307/20528818>
- Forman, E. A., Ramirez-DelToro, V., Brown, L., & Passmore, C. (2017). Discursive strategies that foster an epistemic community for argument in a biology classroom. *Learning and Instruction*, 48, 32–29. <https://doi.org/10.1016/j.learninstruc.2016.08.005>
- Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, 12(3), 23–30. <https://doi.org/10.1111/j.1745-3992.1993.tb00539.x>
- Frey, N., & Fisher, D. (2013). Using common formative assessments as a source of professional development in an urban elementary school. *Teaching and Teacher Education*, 25(5), 674–680. <https://doi.org/10.1016/j.tate.2008.11.006>
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, 33(2), 188–192. <https://doi.org/10.1080/02796015.2004.12086241>
- Fuchs, L. (2016). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research & Practice*, 32(1), 5–7. <https://doi.org/10.1111/ldrp.12127>
- Fuchs, D., & Bradley, R. (2012). A review of Deno and Mirkin's special education resource teacher (SERT) model. In C. A. Espin, K. L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 27–36). University of Minnesota Press.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, 57(6), 488–500. <https://doi.org/10.1177/001440299105700603>

- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53(3), 199–208. <https://doi.org/10.1177/001440298605300301>
- Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children*, 39(5), 14–20. <https://doi.org/10.1177/004005990703900503>
- Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, 73(3), 311–330. <https://doi.org/10.1177/001440290707300303>
- Fuchs, L. S., Fuchs, D., & Hollenbeck, K. N. (2007). Extending responsiveness to intervention to mathematics at first and third grades. *Learning Disabilities Research & Practice*, 22(1), 13–24. <https://doi.org/10.1111/j.1540-5826.2007.00227.x>
- Fuchs, L. S., Fuchs, D., & Stecker, P. M. (1989). Effects of curriculum-based measurement on teachers' instructional planning. *Journal of Learning Disabilities*, 22(1), 51–59. <https://doi.org/10.1177/002221948902200110>
- Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching*, 49(9), 1181–1210. <https://doi.org/10.1002/tea.21054>
- Furtak, E. M., Bakeman, R., & Buell, J. Y. (2018). Developing knowledge-in-action with a learning progression: Sequential analysis of teachers' questions and responses to student ideas. *Teaching and Teacher Education*, 76, 267–282. <https://doi.org/10.1016/j.tate.2018.06.001>
- Furtak, E. M., Cartun, A., Chrzanowski, A., Circi, R., Grover, R., Heredia, S. C., Johnson, R., McClelland, A., & White, K. H. O. (2015, April 16–20). *Toward a participation metaphor for formative assessment* [Paper presentation]. American Educational Research Association Annual Meeting, Chicago, IL, United States.
- Furtak, E. M., Circi, R., & Heredia, S. C. (2018). Exploring alignment among learning progressions, teacher-designed formative assessment tasks, and student growth: Results of a four-year study. *Applied Measurement in Education*, 31(2), 143–156. <https://doi.org/10.1080/08957347.2017.1408624>
- Furtak, E. M., & Heredia, S. C. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching*, 51(8), 982–1020. <https://doi.org/10.1002/tea.21156>
- Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science*, 44, 267–291. <https://doi.org/10.1007/s11251-016-9371-3>
- Furtak, E. M., Morrison, D. L., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education*, 98(4), 640–673. <https://doi.org/10.1002/sce.21122>
- Furtak, E. M., & Ruiz-Primo, M. A. (2008). Making students' thinking explicit in writing & discussion: An analysis of formative assessment prompts. *Science Education*, 92(5), 799–824. <https://doi.org/10.1002/sce.20270>

- Furtak, E. Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education*, 21(4), 360–389. <https://doi.org/10.1080/08957340802347852>
- Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal*, 56(6), 2077–2115. <https://doi.org/10.3102/0002831219843292>
- Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *Elementary School Journal*, 109(5), 537–553. <https://doi.org/10.1086/597001>
- Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education*, 20(2), 150–169. <https://doi.org/10.1080/0969594X.2012.749212>
- Gao, X., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education*, 7(4), 323–342. https://doi.org/10.1207/s15324818ame0704_4
- Gersten, R., Clarke, B. S., Haymond, K., & Jordan, N. C. (2011). Screening for mathematics difficulties in K–3 students. *WestEd Center on Instruction*. <https://files.eric.ed.gov/fulltext/ED524577.pdf>
- Gesel, S. A., LeJeune, L. M., Chow, J. C., Sinclair, A. C., & Lemons, C. J. (2021). A meta-analysis of the impact of professional development on teachers' knowledge, skill, and self-efficacy in data-based decision-making. *Journal of Learning Disabilities*, 54(4), 269–283. <https://doi.org/10.1177/0022219420970196>
- Gitomer, D. H., & Bell, C. A. (2016). Introduction. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (pp. 1–6). AERA.
- Gitomer, D. H., & Duschl, R. A. (1996). Moving toward a portfolio culture in science education. In D. H. Gitomer & R. A. Duschl (Eds.), *Learning science in the schools* (pp. 299–325). Taylor & Francis.
- Gitomer, D. H., & Duschl, R. A. (2007). Indicator systems: establishing multilevel coherence in assessment. *Teachers College Record*, 109(13), 288–320. <https://doi.org/10.1177/016146810710901311>
- Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4), 521–563. <https://doi.org/10.1080/10508406.2013.837391>
- Gochyyev, P., & Wilson, M. (2021, September 26–29). *New curriculum efficacy study and Lord's paradox* [Paper presentation]. Society for Research on Educational Effectiveness Annual Meeting, Arlington, VA, United States. <https://sree.confex.com/sree/2021/meetingapp.cgi/Paper/2754>
- Goertz, M. E., Oláh, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction* (CPRE Research Report No. RR-65).

- Consortium for Policy Research in Education. <https://files.eric.ed.gov/fulltext/ED519791.pdf>
- Goldin, I., Pavlike, P., & Ritter, S. (2016). Discovering domain models in learning curve data. In R. A. Sottolare, A. C. Graesser, X. Hu, A. Olney, B. D. Nye, & A. M. Sinatra (Eds.), *Design recommendations for intelligent tutoring systems: Vol. 4. Domain modeling* (pp. 115–126). U.S. Army Research Laboratory.
- Gong, B. (2021, March 3). *Could two through year assessment designs provide both summative and instructional information: An exploration of feasibility for a through year assessment system*. National Center for the Improvement of Educational Assessment. <https://www.nciea.org/blog/state-testing/could-two-through-year-assessment-designs-provide-both-summative-and>
- Goral, D. P., & Bailey, A. L. (2019). Student self-assessment of oral explanations: Use of language learning progressions. *Language Testing*, 36(3), 391–418. <https://doi.org/10.1177/0265532219826330>
- Graf, E. A., & van Rijn, P. W. (2016). Learning progressions as a guide for design: Recommendations based on observations from a mathematics assessment. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 165–189). Taylor & Francis.
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4), 523–547. <https://www.journals.uchicago.edu/doi/10.1086/681947>
- Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Corwin.
- Hadwin, A. F., Järvelä, S., and Miller, M. (2011). Self-regulated, co-regulated, and socially shared regulation of learning.” In D. Schunk & B. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 65–84). Routledge.
- Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, 55(1), 23–46. <https://doi.org/10.3102/00346543055001023>
- Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, 25(12), 1509–1528. <https://doi.org/10.1080/0950069022000038268>
- Hall, E., Gagnon, D., Schneider, C. M., Thompson, J., & Marion, S. (2014). *State practices related to the use of student achievement measures in the evaluation of teachers in non-tested subjects and grades*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/sites/default/files/publications/Gates_NTGS_Hall_082614.pdf
- Hamilton, A., Wiliam, D., & Hattie, J. (2023, August). The future of AI in education: 13 things we can do to minimize the damage [Working paper]. <https://edarxiv.org/372vr/>
- Hanushek, E. A. (2011). Valuing teachers: How much is a good teacher worth? *Education Next*, 11(3), 41–45. <https://www.educationnext.org/valuing-teachers/>

- Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). The achievement gap fails to close. *Education Next*, 19(3), 8–17. <https://www.educationnext.org/achievement-gap-fails-close-half-century-testing-shows-persistent-divide/>
- Harlen, W. (2005). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 103–117). Sage.
- Harris, L. R., Brown, G. T. L., & Harnett, J. A. (2014). Understanding classroom feedback practices: A study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation, and Accountability*, 26, 107–133. <https://doi.org/10.1007/s11092-013-9187-5>
- Harris, M. J., & Rosenthal, R. (1985). Mediation of interpersonal expectancies effects: 31 meta-analyses. *Psychological Bulletin*, 97, 363–386. <https://doi.org/10.1037/0033-2909.97.3.363>
- Hattie, J. (2009). *Visible learning*. Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (REL 2007–No. 039). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039_sum.pdf
- Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for “Measuring how benchmark assessments affect student achievement”* (REL 2008–No. 002). U.S. Department of Education. https://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/techbrief/tr_00208.pdf
- Heppen, J., Faria, A., Thomsen, K., Sawyer, K., Townsend, M., Kutner, M., Stachel, S., Lewis, S., & Casserly, M. (2010, December). *Using data to improve instruction in the Great City Schools: Key dimensions of practice*. Council of the Great City Schools and American Institutes for Research. https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Strand%202%20Report%20-%20Key%20Dimensions%20of%20Data%20Use_122110.pdf
- Heppen, J., Jones, W., Faria, A., Sawyer, K., Lewis, S., Horowitz, A., Simon, C., Uzzell, R., & Casserly, M. (2011, January). *Using data to improve instruction in the Great City Schools: Documenting current practice*. Council of the Great City Schools and American Institutes for Research. Strand 2 Report - Documenting Current Practice.pdf
- Heredia, S. C., Furtak, E. M., Morrison, D., & Renga, I. P. (2016). Science teachers’ representations of classroom practice in the process of formative assessment design. *Journal of Science Teacher Education*, 27, 697–716. <https://doi.org/10.1007/s10972-016-9482-3>
- Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Council of Chief State School Officers. <https://csaa.wested.org/resource/learning-progressions-supporting-instruction-and-formative-assessment/>
- Heritage, M., & Bailey, A. L. (2006). Assessing to teach: An introduction. *Educational Assessment*, 11(3–4), 145–148. <https://doi.org/10.1080/10627197.2006.9652987>

- Heritage, M., & Heritage, J. (2013). Teacher questioning: The epicenter of instruction and assessment. *Applied Measurement in Education*, 26(3), 176–190. <https://doi.org/10.1080/08957347.2013.793190>
- Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, 28(3), 24–31. <https://doi.org/10.1111/j.1745-3992.2009.00151.x>
- Herman, J., Epstein, S., Leon, S., La Torre Matrundola, D., Reber, S., & Choi, K. (2015). *Implementation and effects of LDC and MDC in Kentucky districts* (CRESST Policy Brief No. 13). University of California, National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/PB_13.pdf
- Hess, K. K., & Kearns, J. (2010, December). *Learning progression frameworks designed for use with the Common Core State Standards in Mathematics K–12*. National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. (updated, v.3) https://www.nciea.org/wp-content/uploads/2022/07/Math_LPF_KH11.pdf
- Hess, K. K., & Kearns, J. (2011, December). *Learning progression frameworks designed for use with the Common Core State Standards in English Language Arts & Literacy K–12*. National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment. http://www.naacpartners.org/publications/ELA_LPF_12.2011_final.pdf
- Higgins, K. M., Harris, N. A., & Kuehn, L. L. (1994). Placing assessment into the hands of young children: A study of student-generated criteria and self-assessment. *Educational Assessment*, 2(4), 309–324. https://doi.org/10.1207/s15326977ea0204_3
- Hosp, J. L., & Hosp, M. (2012). When the emerging alternative became the standard. In C. A. Espin, K. L. McMaster, S. Rose, & M. Miura Wayman (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 49–56). University of Minnesota Press.
- Immekus, J. C., & Atitya, B. (2016). The predictive validity of interim assessment scores based on the full-information bi-factor model for the prediction of end-of-grade test performance. *Educational Assessment*, 21(3), 176–195. <https://doi.org/10.1080/10627197.2016.1202108>
- Ing, M., Chinen, S., Jackson, K., & Smith, T. M. (2021). When should I use a measure to support instructional improvement at scale? The importance of considering both intended and actual use in validity arguments. *Educational Measurement: Issues and Practice*, 40(1), 92–100. <https://doi.org/10.1111/emip.12393>
- James, M. (2017). (Re)viewing assessment: Changing lenses to refocus on learning. *Assessment in Education*, 24(3), 404–414. <https://doi.org/10.1080/0969594X.2017.1318823>
- James, M., & Pedder, D. (2006). Beyond method: Assessment and learning practices and values. *The Curriculum Journal*, 17(2), 109–138. <https://doi.org/10.1080/09585170600792712>
- Jenkins, J., & Fuchs, L. S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In C. A. Espin, K. L. McMaster, S. Rose, & M. Miura Wayman

- (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 7–23). University of Minnesota Press.
- Johnson, C. C., Sondergeld, T. A., & Walton, J. B. (2019). A study of the implementation of formative assessment in three large urban districts. *American Educational Research Journal*, 56(6), 2408–2438. <https://doi.org/10.3102/0002831219842347>
- Jonsson, A. (2013). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education*, 14(1), 63–76. <https://doi.org/10.1177/1469787412467125>
- Jonsson, A., & Panadero, E. (2018). Facilitating students' active engagement with feedback. In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 531–553). Cambridge University Press.
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-analysis. *Learning Disabilities Research & Practice*, 33(3), 144–155. <https://doi.org/10.1111/ldrp.12172>
- Ker, H. W. (2016). The impacts of student-, teacher-and school-level factors on mathematics achievement: an exploratory comparative investigation of Singaporean students and the USA students. *Educational Psychology*, 36(2), 254–276. <https://doi.org/10.1080/01443410.2015.1026801>
- Kingston, N. M., & Broaddus, A. (2017). The use of learning map systems to support the formative assessment in mathematics. *Education Sciences*, 7(1), 41. <https://doi.org/10.3390/educsci7010041>
- Kingston, N. M., Karvonen, M., Bechard, S., & Erickson, K. A. (2016). The philosophical underpinnings and key feathers of the Dynamic Learning Maps alternate assessment. *Teachers College Record*, 118(14), pp. 1–30. <https://doi.org/10.1177/016146811611801410>
- Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37. [Erratum (2015), 34(2), 55.] <https://doi.org/10.1111/j.1745-3992.2011.00220.x>
- Kirschenbaum, H., Napier, R., & Simon, S. B. (1971). *Wad-ja-get? The grading game in American education*. Hart.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>
- Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence* (REL 2017–259). U.S. Department of Education. <https://files.eric.ed.gov/fulltext/ED572929.pdf>
- Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019): A meta-analysis on the impact of grades and comments on

- academic motivation and achievement: A case for written feedback. *Educational Psychology*, 41(7), 922–947. <https://doi.org/10.1080/01443410.2019.1659939>
- Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2018). *Automated student model improvement*. In Proceedings of the 5th International Conference on Educational Data Mining. EDM. <https://files.eric.ed.gov/fulltext/ED537201.pdf>
- Konstantopoulos, S., Li, W., Miller, S. R., & van der Ploeg, A. (2016). Effects of interim assessments across the achievement distribution: Evidence from an experiment. *Educational and Psychological Measurement*, 76(6), 587–608. <https://doi.org/10.1177/0013164415606498>
- Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of diagnostic assessments on mathematics achievement. *Educational Evaluation and Policy Analysis*, 35(4), 481–499. <https://doi.org/10.3102/0162373713498930>
- Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, 9(Suppl. 1), 188–208. <https://doi.org/10.1080/19345747.2015.1116031>
- Kopriva, R. J., Thurlow, M. L., Perie, M., Lazarus, S. S., & Clark, A. (2016). Test takers and the validity of score interpretations. *Educational Psychologist*, 51(1), 108–128. <https://doi.org/10.1080/00461520.2016.1158111>
- Koretz, D. (2017). *The testing charade: Pretending to make schools better*. University of Chicago Press.
- Kroog, H. I., Ruiz-Primo, M. A., & Sands, D. (2014, April 3–7). *Understanding the interplay between the cultural context of classrooms and formative assessment* [Paper presentation]. American Educational Research Association Annual Meeting, Philadelphia, PA, United States.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, 47(2), 211–232. <https://doi.org/10.3102/00346543047002211>
- Lachlan-Hache, L., Matlach, L., Guiden, A., & Castro, M. (2015). *What we know about SLOs: An annotated bibliography of research on and evaluations of student learning objectives*. American Institutes for Research. <https://www.air.org/resource/what-we-know-about-slos-annotated-bibliography-research-and-evaluations-student-learning>
- Lam, E. A., McMaster, K. L., & Rose, S. (2020). Systematic review of curriculum-based measurement with students who are deaf. *The Journal of Deaf Studies and Deaf Education*, 25(4), 398–410. <https://doi.org/10.1093/deafed/ena020>
- Lane, S., & DePascale, C. (2016). *Psychometric considerations for performance-based assessments and student learning objectives. Meeting the challenges to measurement in an era of accountability*. Routledge. <https://doi.org/10.4324/9780203781302>
- Lee, E., & Lee, S. (2009). Effects of instructional rubrics on class engagement behaviors and the achievement of lesson objectives by students with mild mental retardation and their typical peers. *Education and Training in Developmental Disabilities*, 44(3), 396–408. <https://www.jstor.org/stable/24233483>

- Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K–12 education: A systematic review. *Applied Measurement in Education*, 33(2), 124–140. <https://doi.org/10.1080/08957347.2020.1732383>
- Lehrer, R. (2021, August 16–20). *Accountable assessment* [Keynote address]. ACER 2021 Research Conference. <https://research.acer.edu.au/cgi/viewcontent.cgi?article=1009&context=rc21-30>
- Lehrer, R., Kim, M.-J., Ayers, E., & Wilson, M. (2014). Toward establishing a learning progression to support the development of statistical reasoning. In A. Maloney, J. Confrey, & K. Nguyen (Eds.), *Learning over time: Learning trajectories in mathematics education* (pp. 31–60). Information Age Publishers.
- Leighton, J. P. (2019). Cognitive diagnosis is not enough: The challenge of measuring learning with classroom assessments. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 27–45). Routledge.
- Leighton, J. P., Chu, M.-W., & Seitz, P. (2013). Errors in student learning and assessment: The learning errors and formative feedback (LEAFF) model. In R. W. Lissitz (Ed.), *Informing the practice of teaching using formative and interim assessment* (pp. 185–208). Information Age.
- Lekholm, A. K., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: Effects of gender and family background. *Educational Research and Evaluation*, 14(2), 181–199. <https://doi.org/10.1080/13803610801956663>
- Lipnevich, A. A., & Smith, J. K. (2009a). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, 15(4), 319–333. <https://doi.org/10.1037/a0017841>
- Lipnevich, A. A., & Smith, J. K. (2009b). "I really need feedback to learn:" Students' perspectives on the effectiveness of the differential feedback messages. *Educational Assessment, Evaluation and Accountability*, 21, 347–367. <https://doi.org/10.1007/s11092-009-9082-2>
- Liu, L., Hao, J., Andrews, J. J., Zhu, M., Mislevy, R. J., Kyllonen, P., von Davier, A. A., Kerr, D., Ricarte, T., & Graesser, A. (2017). *Collaborative problem solving: Innovating standardized assessment*. International Society of the Learning Sciences. <https://doi.org/10.22318/CSCL2017.118>
- Liu, R., & Koedinger, K. R. (2017). Closing the loop: Automated data-driven cognitive model discoveries lead to improved instruction and learning gains. *Journal of Educational Data Mining*, 9(1), 25–41. <https://doi.org/10.5281/zenodo.3554625>
- Lohbeck, A. (2019). Social and dimensional comparison effects on academic self-concepts and self-perceptions of effort in elementary school children. *Educational Psychology*, 39(1), 133–150. <https://doi.org/10.1080/01443410.2018.1527018>
- Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing*. Information Age Publishing.
- Marion, S. F., & Buckley, K. (2012). *Approaches and considerations for incorporating student performance results from "non-tested" grades and subjects into educator effectiveness determinations*. The National Center for the Improvement of Educational

- Assessment. https://www.nciea.org/sites/default/files/publications/Marion_Buckley_Considerations_for_non-tested_grades_2011.pdf
- Marion, S. F., DePascale, C., Domaleski, C., Gong, B., & Diaz-Bilello, E. (2012). *Considerations for analyzing educators' contributions to student learning in non-tested subjects and grades with a focus on student learning objectives*. The National Center for the Improvement of Educational Assessment. https://www.nciea.org/sites/default/files/publications/Measurement%20Considerations%20for%20NTSG_052212%20v2.pdf
- Marsh, J. A. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record*, 114(11), 1–48. <https://doi.org/10.1177/016146811211401106>
- Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practices. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129–185). Buros Institute of Mental Measurements, University of Nebraska.
- Marston, D. (2012). School- and district-wide implementation of curriculum-based measurement in the Minneapolis public schools. In C. A. Espin, K. L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 59–78). University of Minnesota Press.
- Martone, A., Reagan, D., & Reed, G. (2018). Understanding the use of mathematics interim assessments: A case study. *International Electronic Journal of Elementary Education*, 10(5), 515–523. <https://www.iejee.com/index.php/IEJEE/article/view/318>
- Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. University of Nebraska–Lincoln. <http://dwb.unl.edu/Edit/MB/MasonBruning.html>
- Massachusetts Department of Education. (1999). *1999 MCAS technical report*. https://archive.org/details/ERIC_ED459184/page/n1
- McBee, M. M., & Barnes, L. L. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education*, 11(2), 179–194. https://doi.org/10.1207/s15324818ame1102_4
- McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education*, 41(2), 68–84. <https://doi.org/10.1177/00224669070410020301>
- McMaster, K. L., Ritchey, K. D., & Lembke, E. (2011). Curriculum-based measurement for beginning writers: Recent developments and future directions. *Assessment and Intervention*, 24, 111–148. [https://doi.org/10.1108/S0735-004X\(2011\)0000024008](https://doi.org/10.1108/S0735-004X(2011)0000024008)
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1), 20–32. <https://doi.org/10.1111/j.1745-3992.2001.tb00055.x>
- McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research*, 95(4), 203–213. <https://doi.org/10.1080/00220670209596593>

- Michaels, S., Shouse, A. W., & Schweingruber, H. A. (2008). *Ready, set, science!* National Academies Press.
- Michigan Assessment Consortium. (2017, Fall). *Assessment literacy standards*. Michigan Assessment Consortium. <https://www.michiganassessmentconsortium.org/assessment-literacy-standards/>
- Miner, B. C. (1967). Three factors of school achievement. *The Journal of Educational Research*, 60(8), 370–376. <https://doi.org/10.1080/00220671.1967.10883518>
- Minstrell, J., Anderson, R., & Li, M. (2010). *Assessing teacher competency in formative assessment: Annual report 2010*. National Science Foundation.
- Mislevy, R. J. (2018a). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test data analytics and psychometrics: Informing assessment practices* (pp. 1–52). Information Age.
- Mislevy, R. J. (2018b). *Sociocognitive foundations of educational measurement*. Routledge.
- Missouri Department of Education. (2015). *Setting growth targets for student learning objectives: Methods and considerations*. <https://dese.mo.gov/sites/default/files/Methods-and-Considerations.pdf>
- Miura Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41(2), 85–120. <https://doi.org/10.1177/00224669070410020401>
- Moore, C. C. (1939). The elementary school mark. *The Pedagogical Seminary and Journal of Genetic Psychology*, 54, 285–294.
- Morell, L., Collier, T., Black, P., & Wilson, M. (2017). A construct-modeling approach to develop a learning progression of how students understand the structure of matter. *Journal of Research in Science Teaching*, 54(8), 1024–1048. <https://doi.org/10.1002/tea.21397>
- Morton, C. (2013). *Judging alignment of curriculum-based measures in mathematics and common core standards* [Unpublished doctoral dissertation, University of Oregon]. https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/17879/Morton_oregon_0171A_10841.pdf?sequence=1&isAllowed=y
- Mosher, F. A. (2022). Learning progressions. In D. Fisher (Ed.), *Routledge resources online: Education*. <https://doi.org/10.4324/9781138609877-REE115-1>
- Myers, C. B. (1996, April 8–12). *Beyond the PDS: Schools as professional learning communities: A proposal based on an analysis of PDS efforts of the 1990s* [Paper presentation]. American Educational Research Association Annual Meeting, New York, NY, United States.
- National Center on Response to Intervention. (March 2010). *Essential components of RTI—a closer look at Response to Intervention*. U.S. Department of Education. <https://files.eric.ed.gov/fulltext/ED526858.pdf>
- National Centre for Excellence in the Teaching of Mathematics. (2020). *Mathematics guidance: Key stages 1 and 2: Non-statutory guidance for the national curriculum in England*. UK Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1017683/Maths_guidance_KS_1_and_2.pdf

- National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K–8*. National Academies Press. <https://nap.nationalacademies.org/read/11625/chapter/1>
- Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155–175. https://doi.org/10.1207/s15326985ep2202_4
- New York State Education Department. (2014, March). *Guidance on the New York State district-wide growth goal-setting process for teachers: Student learning objectives*.
- Newton, P. E. (2010). The multiple purposes of assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed, pp. 392–396). Elsevier.
- No Child Left Behind Act of 2001, P.L. 107–110, 20 U.S.C. § 6319 (2002).
- Northwest Evaluation Association. (2019, March). *MAP® Growth™ technical report*.
- Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, 85(2), 226–245. <https://doi.org/10.1080/01619561003688688>
- Olsen, B., & Buchanan, R. (2019). An investigation of teachers encouraged to reform grading practices in secondary schools. *American Educational Research Journal*, 56(5), 2004–2039. <https://doi.org/10.3102/0002831219841349>
- Opper, I. M. (2019). *Teachers matter: Understanding teachers' impact on student achievement*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR4312.html
- Otero, V. K. (2006). Moving beyond the “get it or don’t” conception of formative assessment. *Journal of Teacher Education*, 57(3), 247–255. <https://doi.org/10.1177/0022487105285963>
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129–144. <https://doi.org/10.1016/j.edurev.2013.01.002>
- Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, 22(6), 806–813. <https://doi.org/10.1016/j.lindif.2012.04.007>
- Parsi, A., & Darling-Hammond, L. (2015). *Performance assessments: How state policy can advance assessments for 21st century learning*. National Association of State Boards of Education, Stanford Center for Opportunity Policy in Education. <https://pdfs.semanticscholar.org/d479/05d18b23dd7eb367c0627656e3e7f321567b.pdf>
- Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. National Board on Educational Testing and Public Policy. <https://www.bc.edu/research/nbetpp/statements/nbr2.pdf>
- Pellegrino, J. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa*, 20(2), 65–77. <https://doi.org/10.1016/j.pse.2014.11.002>

- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. <https://doi.org/10.1080/00461520.2016.1145550>
- Penuel, W. R., & Shepard, L. A. (2016). Assessment and teaching. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 787–850). American Educational Research Association.
- Pereira, M., & Tienken, C. (2012). An evaluation of the influence of interim assessments on Grade 8 student achievement in mathematics and language arts. *International Journal of Educational Leadership Preparation, 7*(3), 1–13. <https://files.eric.ed.gov/fulltext/EJ997471.pdf>
- Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13. <https://doi.org/10.1111/j.1745-3992.2009.00149.x>
- Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes: Towards a wider conceptual field. *Assessment in Education, 5*(1), 85–102. <https://doi.org/10.1080/0969595980050105>
- Peters, R., Kruse, J., Buckmiller, T., & Townsley, M. (2017). “It’s just not fair!” Making sense of secondary students’ resistance to a standards-based grading. *American Secondary Education, 45*(3), 9–28. <https://www.proquest.com/docview/1938071902>
- Peterson, P. (1979). Direct instruction reconsidered. In P. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 57–65). McCutchan.
- Pham, D. N., Wells, C. S., Bauer, M. I., Wylie, C., & Monroe, S. (2021). Examining three learning progressions in middle-school mathematics for formative assessment. *Applied Measurement in Education, 34*(2), 107–121. <https://doi.org/10.1080/08957347.2021.1890744>
- Pitt, E., & Norton, L. (2017). “Now that’s the feedback I want!” Students’ reactions to feedback on graded work and what they do with it. *Assessment and Evaluation in Higher Education, 42*(4), 499–516. <https://doi.org/10.1080/02602938.2016.1142500>
- Pokorny, H., & Pickford, P. (2010). Complexity, cues, and relationships: Student perceptions of feedback. *Active Learning in Higher Education, 11*(1), 21–30. <https://doi.org/10.1177/1469787409355872>
- Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record, 117*(11), 1–28. <https://doi.org/10.1177/016146811511701106>
- Porter, A. C. (1993). School delivery standards. *Educational Researcher, 22*(5), 24–30. <https://doi.org/10.3102/0013189X022005024>
- Powell, D., Lamba, S., Ismail, K., & Marland, J. (2022). *What are through-year assessment? Exploring multiple approaches to through-year design*. Education First. <https://www.education-first.com/wp-content/uploads/2023/01/What-are-Through-year-Assessments-1.pdf>

- Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement*, 46(1), 1–18. <https://doi.org/10.1111/j.1745-3984.2009.01066.x>
- Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, 26(7), 1372–1380. <https://doi.org/10.1016/j.tate.2010.03.008>
- Randel, B., Apthorp, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *Journal of Educational Research*, 109(5), 491–502. <https://doi.org/10.1080/00220671.2014.992581>
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435–448. <https://doi.org/10.1080/02602930902862859>
- Resnick, L. B., & Resnick, D. P. (1992) Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Evaluation in education and human services* (Vol. 30, pp. 37–75). Springer. https://doi.org/10.1007/978-94-011-2968-8_3
- Rhoades, K., & Madaus, G. (2003). *Errors in standardized tests: A systemic problem*. National Board on Educational Testing and Public Policy. <https://files.eric.ed.gov/fulltext/ED479797.pdf>
- Riggan, M., & Oláh, L. N. (2011). Locating interim assessments within teachers' assessment practice. *Educational Assessment*, 16(1), 1–14. <https://doi.org/10.1080/10627197.2011.551085>
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255. <https://doi.org/10.3758/BF03194060>
- Rosenshine, B. V. (1979). Content, time, and direct instruction. In P. Peterson & H. Walberg (Eds.), *Research on teaching: Concepts, findings, and implications* (pp. 28–56). McCutchan.
- Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in Grade 5–6 mathematics: Effects on problem-solving achievement. *Educational Assessment*, 8(1), 43–58. https://doi.org/10.1207/S15326977EA0801_03
- Ross, J. A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of Grade 9 geography. *Assessment in Education*, 15(2), 183–199. <https://doi.org/10.1080/09695940802164218>
- Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. Jossey-Bass.
- Rowe, K. J., & Hill, P. W. (1996) Assessing, recording and reporting students' educational progress: The case for "subject profiles." *Assessment in Education*, 3(3), 309–352. <https://doi.org/10.1080/0969594960030304>
- Ruiz-Primo, M. A., & Brookhart, S. M. (2018). *Using feedback to improve learning*. Routledge.

- Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3–4), 205–235. <https://doi.org/10.1080/10627197.2006.9652991>
- Ruiz-Primo, M. A., & Furtak, E. M. (2007). Informal formative assessment and scientific inquiry: Exploring teachers' practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84. <https://doi.org/10.1002/tea.20163>
- Ruiz-Primo, M. A., Kroog, H. I., & Sands, D. I. (2015, August 25–29). *Teacher judgments on-the-fly: Teachers' response patterns in the context of informal formative assessment* [Paper presentation]. EARLI Biennial Conference, Limassol, Cyprus.
- Ruiz-Primo, M. A., & Li, M. (2013). Analyzing teachers' feedback practices in response to students' work in science classrooms. *Applied Measurement in Education, 26*(3), 163–175. <https://doi.org/10.1080/08957347.2013.793188>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. <https://doi.org/10.1007/BF00117714>
- Sadler, D. R. (2014). The futility of attempting to codify academic achievement standards. *Higher Education, 67*(3), 273–288. <https://doi.org/10.1007/s10734-013-9649-1>
- Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066. <https://doi.org/10.1037/edu0000190>
- Scalise, K., & Wilson, M. (2011). The nature of assessment systems to support effective use of evidence through technology. *E-Learning and Digital Media, 8*(2), 121–132. <https://doi.org/10.2304/elea.2011.8.2.121>
- Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Applied Measurement in Education, 26*(3), 159–162. <https://doi.org/10.1080/08957347.2013.793189>
- Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education, 26*(3), 191–204. <https://doi.org/10.1080/08957347.2013.793185>
- Schneider, M. C., & Johnson, R. L. (2018). *Using formative assessment to support student learning objectives*. Routledge.
- Schwartz, R., Ayers, E., & Wilson, M. (2017). Mapping a data modeling and statistical reasoning learning progression using unidimensional and multidimensional item response models. *Journal of Applied Measurement, 18*(3), 268–298.
- Shavelson, R. J. (2009, June 24–26). *Reflections on learning progressions* [Paper presentation]. Learning Progressions in Science Conference, Iowa City, IA, United States.
- Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295–314. <https://doi.org/10.1080/08957340802347647>

- Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20(7), 2–16. <https://doi.org/10.3102/0013189X020007002>
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14. <https://doi.org/10.3102/0013189X029007004>
- Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, 85(2), 246–257. <https://doi.org/10.1080/01619561003708445>
- Shepard, L. A. (2018). Learning progressions as tools for assessment and learning. *Applied Measurement in Education*, 31(2), 165–174. <https://doi.org/10.1080/08957347.2017.1408628>
- Shepard, L. A., Penuel, W. R., & Davidson, K. L. (2017). Design principles for new systems of assessment. *Phi Delta Kappan*, 98(6), 47–52. <https://doi.org/10.1177/0031721717696478>
- Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, 73, 131–149. <https://doi.org/10.1016/j.jsp.2019.03.005>
- Shriner, J. G., & Thurlow, M. L. (2012). Curriculum-based measurement, progress monitoring, and state assessments. In C.A. Espin, K.L. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education*, (pp. 247–260). University of Minnesota Press.
- Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14. <https://doi.org/10.3102/0013189X015002004>
- Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Simon, M. A. (1995) Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education*, 26(2), 114–145. <https://doi.org/10.5951/jresmetheduc.26.2.0114>
- Smarter Balanced Assessment Consortium. (2019). 2020–2021 Interim assessments overview. <https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf>
- Smith, A. Z., & Dobbin, J. E. (1960). Marks and marking systems. In C. W. Harris (Ed.), *Encyclopedia of educational research* (3rd ed., pp. 783–791). Macmillan.
- Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. <https://doi.org/10.1177/1745691615569000>
- Solberg, L. I., Mosser, G., & McDonald, S. (1997). The three faces of performance measurement: improvement, accountability, and research. *The Joint Commission journal on quality improvement*, 23(3), 135–147. [https://doi.org/10.1016/s1070-3241\(16\)30305-4](https://doi.org/10.1016/s1070-3241(16)30305-4)
- Starch, D. (1913). Reliability and distribution of grades. *Science*, 38(983), 630–636. <https://doi.org/10.1126/science.38.983.630>
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, 42(8), 795–819. <https://doi.org/10.1002/pits.20113>

- Stiggins, R. J. (1994). *Student-centered classroom assessment*. Merrill.
- Stiggins, R. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324–328. <https://doi.org/10.1177/003172170508700414>
- Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice*, 8(2), 5–14. <https://doi.org/10.1111/j.1745-3992.1989.tb00315.x>
- Stoeffler, K., Rosen, Y., Bolsinova, M., & von Davier, A. A. (2020). Gamified performance of collaborative problem solving skills. *Computers in Human Behavior*, 104(106036). <https://doi.org/10.1016/j.chb.2019.05.033>
- Takahashi, S., Jackson, K., Norman, J. R., Ing, M., & Krumm, A. E. (2022). Measurement for improvement. In D. Peurach, J. Russell, L. Cohen-Vogel, & W. R. Penuel (Eds.), *The foundational handbook on improvement research in education* (pp. 423–442). Rowman & Littlefield.
- Taylor, J. J., Buckley, K., Hamilton L. S., Stecher, B. M., Read, L., & Schweig, J. (2018). *Choosing and using SEL competency assessments: What schools and districts need to know*. Collaborative for Academic, Social, and Emotional Learning. <http://measuringsel.casel.org/pdf/>
- Templin, J., & Bradshaw, L. (2014). The use and misuse of psychometric models. *Psychometrika*, 79(2), 347–354. <https://doi.org/10.1007/S11336-013-9364-Y>
- Texas Education Agency. (2021, July). *Student learning objectives implementation guide for administrators*. https://texasslo.org/Resource_files/resources/SLO_Administrator_Guide_w_TIA_Tips.pdf
- Thomas, S., & Oldfather, P. (1997). Intrinsic motivations, literacy, and assessment practices: “That’s my grade. That’s me.” *Educational Psychologist*, 32(2), 107–123. https://doi.org/10.1207/s15326985sep3202_5
- Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education*, 2013(958530), 1–29. <http://dx.doi.org/10.1155/2013/958530>
- Topping, K. J. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 395–412). Sage.
- Tunstall, P., & Gipps, C., (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal*, 22(4), 389–404. <https://doi.org/10.1080/0141192960220402>
- Tyler, J., & McNamara, C. (2011, Fall). *An examination of teacher use of the data dashboard student information system in Cincinnati Public Schools*, Vol. VI. Council of the Great City Schools. <https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/85/SUERF%20VI.pdf>
- U.S. Department of Education. (2015, October). *Testing action plan*. <https://www.ed.gov/news/press-releases/fact-sheet-testing-action-plan>

- U.S. Department of Education, National Assessment of Educational Progress. (2020). *1990–2019 mathematics and reading assessments*. https://www.nationsreportcard.gov/dashboards/achievement_gaps.aspx
- U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2009). *Implementing data-informed decision making in schools: Teacher access, supports and use*. <https://www2.ed.gov/about/offices/list/oepdp/ppss/reports.html>
- U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. <https://www2.ed.gov/rschstat/eval/data-to-inform-instruction/report.pdf>
- van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice*, 32(1), 46–60. <https://doi.org/10.1111/ldrp.12122>
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- Van der Kleij, F. M., & Lipnevich, A. A. (2021). Student perceptions of assessment feedback: A critical scoping review and call for research. *Educational Assessment, Evaluation and Accountability*, 33(2), 345–373. <https://doi.org/10.1007/s11092-020-09331-x>
- Vanlommel, K., & Schildkamp, K. (2019). How do teachers make sense of data in the context of high-stakes decision making? *American Educational Research Journal*, 56(3), 792–821. <https://doi.org/10.3102/0002831218803891>
- van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20(4), 270–279. <https://doi.org/10.1016/j.learninstruc.2009.08.004>
- von Davier, A. A. (2017). Computational psychometrics in support of collaborative assessments. *Journal of Educational Measurement*, 54(1), 3–11. <https://doi.org/10.1111/jedm.12129>
- von Davier, A. A., DiCerbo, K., & Verhagen, J. (2021). Computational psychometrics: A framework for estimating learners' knowledge, skills and abilities from learning and assessments systems. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 25–43). Springer.
- von Davier, A. A., Wong, P., Polyak, S. T., & Yudelson, M. (2019). The argument for a “Data Cube” for large-scale psychometric data. *Frontiers in Education*, 4(71). <http://doi.org/10.3389/educ.2019.00071>
- Waltman, K. K., & Frisbie, D. A. (1994). Parents' understanding of their children's report card grades. *Applied Measurement in Education*, 7(3), 223–240. https://doi.org/10.1207/s15324818ame0703_5

- Welsh, M. E., D'Agostino, J. V., & Kaniskan, R. (2013). Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues and Practice*, 32(2), 26–36. <https://doi.org/10.1111/emip.12009>
- Westrick, P. A. (2017). Reliability estimates for undergraduate grade point average. *Educational Assessment*, 22(4), 231–252. <https://doi.org/10.1080/10627197.2017.1381554>
- Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment*, 20(1), 23–45. <https://doi.org/10.1080/10627197.2015.997614>
- Wiggins, G. (1998). *Educative assessment: Designing assessment to inform and improve student performance*. Jossey-Bass..
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18–40). Routledge.
- Wiliam, D. (2018a). *Embedded formative assessment* (2nd ed.). Solution Tree.
- Wiliam, D. (2018b). How can assessment support learning? A response to Wilson and Shepard, Penuel, and Pellegrino. *Educational Measurement: Issues and Practice*, 37(1), 42–44. <https://doi.org/10.1111/emip.12192>
- Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education*, 11(1), 49–65. <https://doi.org/10.1080/0969594042000208994>
- Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Routledge. <https://doi.org/10.4324/9781315086545>
- Wilkerson, S. B., Klute, M., Peery, B., & Liu, J. (2021). *How Nebraska teachers use and perceive summative, interim, and formative data* (REL 2021–054). U.S. Department of Education. <https://files.eric.ed.gov/fulltext/ED610179.pdf>
- Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39(1), 1–37. <https://doi.org/10.1111/j.1745-3984.2002.tb01133.x>
- Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716–730. <https://doi.org/10.1002/tea.20318>
- Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science* (pp. 317–343). Sense Publishers.
- Wilson, M. (2021, August 16–20). *Rethinking measurement for accountable assessment* [Keynote address]. ACER 2021 Research Conference. <https://research.acer.edu.au/cgi/viewcontent.cgi?article=1015&context=rc21-30>
- Wilson, M. (2023). Finding the right grain-size for measurement in the classroom. *Journal of Education and Behavioral Statistics*, 49(1), 3–31. <https://doi.org/10.3102/10769986231159006>

- Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education, Part II* (pp. 132–154). University of Chicago Press.
- Wilson, M., Gochyyev, P., & Scalise, K. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning*, 20(2), 97–119. <https://olj.onlinelearningconsortium.org/index.php/olj/article/view/799/205>
- Wilson, M., & Lehrer, R. (2021). Improving learning: Using a learning progression to coordinate instruction and assessment. *Frontiers in Education*, 6, 654212. <https://doi.org/10.3389/feduc.2021.654212>
- Wilson, M., & Scalise, K. (2016). Learning analytics: Negotiating the intersection of measurement technology and information technology. In J. M. Spector, B. B. Lockee, & M. D. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (pp. 1–23). Springer.
- Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13(2), 181–208. https://doi.org/10.1207/S15324818AME1302_4
- Winstone, N. E., Mathlin, G., & Nash, R. A. (2019). Building feedback literacy: Students' perceptions of the Developing Engagement with Feedback Toolkit. *Frontiers in Education*, 4(39). <https://doi.org/10.3389/feduc.2019.00039>
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017) Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1), 17–37. <https://doi.org/10.1080/00461520.2016.1207538>
- Wong J., Baars, M., de Koning, B. B., van der Zee, T., Davis, D., Khalil, M., Houben, G.-J., & Paas, F. (2019) Educational theories and learning analytics: From data to knowledge. In D. Ifenthaler, D. K. Mah, & J. K. Yau (Eds.), *Utilizing learning analytics to support study success* (pp. 3–25). Springer.
- Wylie, E. D. (2008). *Formative assessment: Examples of practice*. Council of Chief State School Officers. https://www.ccsso.org/sites/default/files/2017-12/Formative_Assessment_Examples_2008.pdf
- Wylie, E. C., & Lyon, C. J. (2015). The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education*, 22(1), 140–160. <https://doi.org/10.1080/0969594X.2014.990416>
- Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education*, 28(3), 228–260. <https://doi.org/10.1080/0969594X.2021.1884042>
- Yao, S.-Y., Wilson, M., Henderson, J. B., & Osborne, J. (2015). Investigating the function of content and argumentation items in a science test: A multidimensional approach. *Journal of Applied Measurement*, 16(2), 171–192.

- Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., & Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573, 364–369. <https://doi.org/10.1038/s41586-019-1466-y>
- Zeng, W., Huang, F., Yu, L., & Chen, S. (2018). Towards a learning-oriented assessment to improve students' learning—a critical review of literature. *Educational Assessment, Evaluation and Accountability*, 30, 211–250. <https://doi.org/10.1007/s11092-018-9281-9>
- Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. R. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, 6(2), 153–174. <http://dx.doi.org/10.18608/jla.2019.62.11>
- Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulated learning and performance* (pp. 49–64). Routledge.

Test-Based Accountability in K–12 Education

Andrew D. Ho

Harvard Graduate School of Education

Morgan S. Polikoff

University of Southern California

Over the past several decades, K–12 educational tests have served a variety of roles in formal and informal accountability policies at the student, teacher, school, district, state, and federal levels. The use of tests for accountability places a considerable burden on those seeking validity evidence (see Lane & Marion, this volume, for a discussion of validity and validation). While a canonical user of a test score may draw an inference about a single student, accountability systems often incorporate many actors, and many derived scores, at many levels of an educational system. Who is holding whom accountable? By what mechanism? Using which scores? For what purpose?

In the face of such complexity, the measurement field has sometimes treated accountability as a distinct second-stage consideration. A developer may first gather validity evidence for a student score interpretation and then leave those who wish to use scores in an accountability system to gather validity evidence for such uses. This two-stage formulation risks isolating measurement experts and measurement expertise in the first stage, when recent history has shown that the second stage often follows the first. A modern approach to analyzing tests from a measurement perspective is integrated. It anticipates the likely use of scores for accountability and aims to document and prevent the bias, variance, and unintended consequences that such uses can impart.

Our chapter extends the significant contribution by Koretz and Hamilton (2006) in the fourth edition of *Educational Measurement*. The first three editions of *Educational Measurement* had no stand-alone chapter on accountability. Koretz and Hamilton provided an extensive treatment of the validity of inferences under high-stakes conditions. Their treatment, further developed by Koretz (2008), convincingly demonstrates how validity evidence gathered under low-stakes administration conditions is inadequate to support the use of tests for accountability. We extend their framework and principles to the wide range of derived scores common in modern accountability systems.

Koretz and Hamilton (2006) also provided a comprehensive history of large-scale testing in the United States, including the increasing use of assessments for accountability from the 1980s through the initial years following the No Child Left Behind Act of 2001 (NCLB). We summarize and extend their history through three additional discernible national accountability policy periods, the Growth Model Pilot Program (GMPP) of 2005, the “waiver” period beginning in 2011, and the Every Student Succeeds Act of 2015 (ESSA). We emphasize how each period brought accompanying measurement challenges, including the measurement of growth, the measurement of college and career readiness, and score comparability across states and over time.

The COVID-19 pandemic closed U.S. schools in early 2020 and disrupted assessment and accountability systems worldwide. The pandemic coincided with the development of this chapter, and our perspective on pandemic-era accountability is limited. While we treat this recent period simplistically as a pause and return of ESSA accountability policies, we acknowledge that states varied considerably in their responses to the pandemic, and current accountability policies are even more varied across states than they were prior to the pandemic. We look forward to future research that summarizes the history and impact of this period of volatility in educational accountability policies.

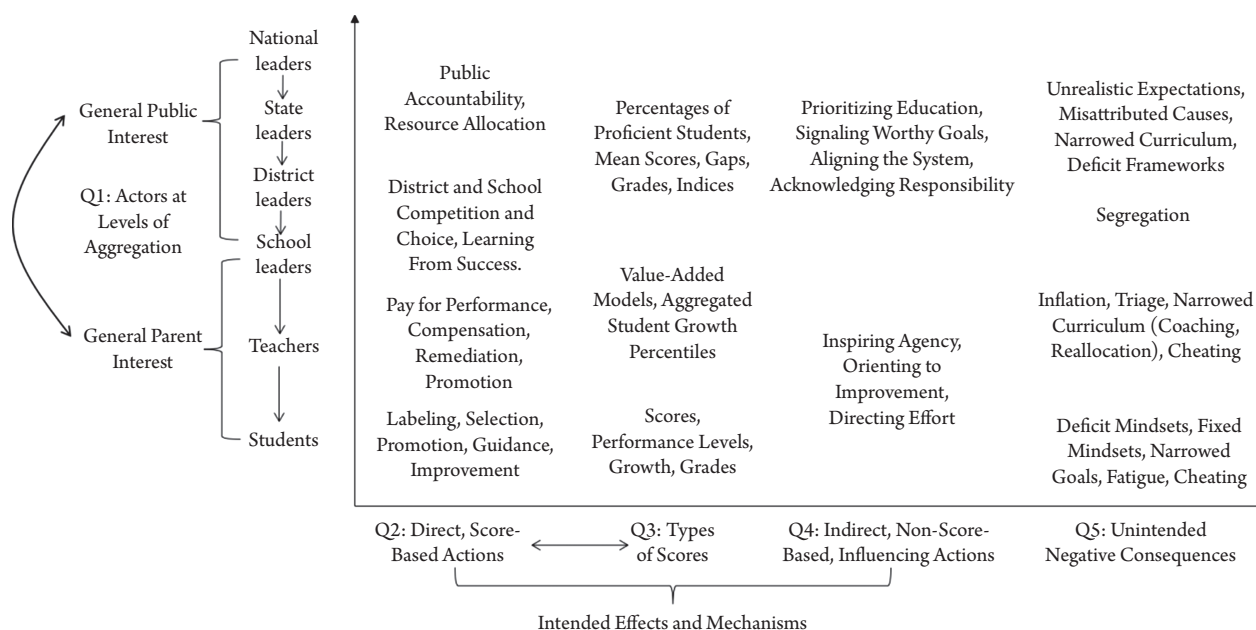
Our chapter complements others in this volume. Our focus is on assessment for accountability in the United States, including the low-stakes, but not no-stakes, National Assessment of Educational Progress (NAEP). In contrast, Braun and Kirsch (this volume) focus on large-scale assessments in other countries, including cross-national monitoring assessments like the Programme for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS). Our focus is on assessment for accountability across multiple levels, including the nation, state, district, school, classroom, teacher, and student levels. Brookhart and DePascale (this volume) cover student-level assessment for purposes like orienting diagnostic feedback and informing pedagogical decisions, and Margolis et al. (this volume) cover assessment for licensing and certification. Our focus is also on school environments in K–12 grade levels. Camara et al. (this volume) cover assessment in the context of higher education. Finally, our focus is on cognitive constructs related to academic content, whereas Kyllonen and Zu (this volume) cover issues related to assessments of social and emotional learning and nonacademic outcomes.

THE LOGIC AND HISTORICAL DEVELOPMENT OF K–12 TEST-BASED ACCOUNTABILITY

Accountability testing uses formal or informal policies and mechanisms to encourage or require actors to use test scores. Five questions can help to distinguish among different approaches to “assessment for accountability”: (1) Who is holding whom accountable? (2) By what mechanism? (3) Using which scores? (4) For what purpose? (5) With what unintended negative consequences? Figure 16.1 provides an overview of these questions asked and answered at different levels of K–12 educational systems. We explain the vertical and horizontal dimensions of Figure 16.1 here and return to this framework throughout the chapter.

The vertical dimension distinguishes among levels of the educational system, answering Question 1, “Who is holding whom accountable?” At higher levels of aggregation, national, state, and district leaders hold lower level leaders accountable to educational progress in the interest of the general public. At lower levels of aggregation, schools and teachers use tests to hold students accountable to learning. We reference “general public interest” and “general parent interest” at higher and lower levels of the educational system, respectively, to indicate that justifications for accountability emphasize more benefits to the general public at higher levels and more benefits for parents and their children in specific classrooms and schools at lower levels. These interests also interact, where parents serve in general public roles and can be a strong advocacy group at higher levels of educational systems.

The horizontal dimension of Figure 16.1 contains four columns, each of which addresses the remaining four questions in turn, about accountability mechanisms, score types, purposes, and unintended negative consequences, respectively. The first column lists accountability decisions and mechanisms that test scores serve at different levels

**FIGURE 16.1****A Framework for K–12 Test-Based Accountability Purposes and Mechanisms**

of the system. For example, at higher levels of aggregation, publicly reported aggregate scores for the nation, states, districts, and schools can act as a form of “public accountability” (Hutt & Polikoff, 2020), where transparency about aggregate scores and progress can theoretically inform the general public and inspire educational improvement. Policies can also establish funding or program enrollment for higher scoring or lower scoring schools, teachers, and students. These policies have higher or lower stakes depending on the consequence of the policy, the individuals or groups it targets, and the relative weight of test scores in the policy decision.

The second column describes the types of scores that usually support these decisions and mechanisms. Emphasizing score types and uses is particularly essential for accountability testing because intended uses are often at different levels of aggregation, or scores are “adjusted” or “combined” in various ways to serve accountability purposes. These “derived scores” have properties that can both serve intended purposes and create unintended consequences (Haertel & Ho, 2016). Chapter 13 of the *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association [AERA] et al., 2014) also discusses these issues in the context of “accountability indices.”

The third column of Figure 16.1 describes “indirect” or “influencing” purposes of accountability testing that also serve as broad justifications for accountability. Haertel (2013) distinguished these indirect purposes of testing as mechanisms, including incentives and messaging, that influence actors without directly using their test scores. For example, a goal of test-based graduation policies is to direct student and teacher effort, and a goal of school accountability policies is to signal worthy curricular goals

and levels of proficiency. Both intended effects occur as the result of the policy, not as a result of interpretations or uses of any particular scores themselves. Indirect purposes are usually but not necessarily explicit in the logic of test-based accountability policies. Validity evidence supporting indirect test uses includes quasi-experimental analyses that estimate the effect of the test-based accountability policy over a counterfactual policy with less testing or accountability.

The final column of Figure 16.1 lists common unintended negative consequences of test-based accountability policies at different levels of educational systems. These can include common responses like inflation that derive from incomplete or unrealistic goal structures (Koretz, 2008). Additional unintended consequences of accountability testing include “deficit interpretations” of students or groups where users interpret low scores or averages as irreparable or innate student or group deficits for themselves or for others. For example, recent empirical evidence suggests that “achievement gap” framing affects teacher support for policy solutions and general public beliefs about racial stereotypes and causes of disparities (Quinn, 2020; Quinn et al., 2019). Professional testing standards are explicit about the importance of anticipating, documenting, and minimizing unintended negative consequences in test-based accountability (AERA et al., 2014).

Avoiding unintended consequences requires resolving a central tension in test-based accountability: on the one hand, between incentivizing achievement of students, teachers, and school leaders to meet test-based goals and, on the other hand, formalizing the responsibility of school, district, state, and federal actors to provide the capacity and opportunity for this achievement. Elmore (2004) laid out five guiding principles for designing fair accountability policies that attempt to resolve this tension: (a) there should be empirical evidence that goals are achievable, (b) measures should be sufficiently accurate and precise for their purposes, (c) students should be held accountable only to what they have had the opportunity to learn, (d) schools should be held accountable only to the value they add, and (e) the performance that a policy requires should be reciprocated by the responsibility of the system to provide that capacity.

Implementing these principles is challenging in part because test-based measures cannot on their own distinguish between a low score that reflects a lack of individual commitment to achievement and a low score that reflects a lack of systematic support for achievement. A lack of individual commitment might suggest that stakes should be higher, whereas a lack of systematic support suggests that support should be stronger. This fundamental ambiguity in test score interpretation enables actors at different levels of the accountability system to use the same test score data to ask more from each other. Higher level actors use policy and authority to require more of lower level actors. Lower level actors can use mechanisms of public accountability (Hutt & Polikoff, 2020) to require more assistance or support from higher level actors.

These mechanisms are not equal in power. Because of this asymmetry, particular care is required to ensure that the system is beneficent in providing opportunities to accelerate learning without reinforcing existing inequalities and power structures. Toward

the end of this chapter, we review recent movements that attempt to rebalance power toward holding systems accountable to learners, including systems of equity indicators (National Academies of Sciences, Engineering, and Medicine [NASEM], 2019), sociocultural assessment (Shepard, 2021), and indigenous data sovereignty and policy (Walter et al., 2021).

Balance between stakes and supports among actors, mechanisms, scores, and purposes has evolved through the history of U.S. federal educational policy. In the remainder of this section, we review five phases in rough chronological order: the Tyler rationale, program evaluation and monitoring, programmed instruction, minimum competency testing, and the standards movement. Our review draws from similar reviews of history relevant to accountability testing, including those by Haertel and Herman (2005), Koretz and Hamilton (2006), and Madaus et al. (1983). Like Shepard et al. (in press), we also update this history by describing recent waves of federal accountability policy, including NCLB and ESSA.

The Tyler Rationale

The Tyler rationale (Tyler, 1949) forms the basis of many modern models for improving curriculum and instruction. Tyler drew lessons from the “Eight-Year Study” (E. Smith & Tyler, 1942) conducted from 1933 to 1941. The original study investigated whether students in 30 high schools with relaxed college-preparatory course requirements had different outcomes than students in high schools that maintained the strict curricular sequences of that era.

Tyler (1949) distilled four principles from the study: (a) define appropriate objectives, (b) design educational experiences aligned to objectives, (c) organize educational experiences for effectiveness, and (d) evaluate whether objectives have been attained. As Haertel and Herman (2005) noted, Tyler and his colleagues also articulated clearly that tests served multiple purposes beyond evaluation. These purposes included focusing student and teacher effort and providing “a sound basis for public relations” (E. Smith & Tyler, 1942, p. 10) through evidence about outcomes. The use of tests to focus the system and shape public perceptions would grow in the decades to come (Haertel, 2013).

Programmed Instruction and Criterion-Referenced Testing

From the 1950s through the 1970s, Tyler’s principles undergirded the rise of programmed instruction, a model that shifted Tyler’s emphasis on measurable objectives from the program level to the classroom level. Psychologists at the time envisioned a well-crafted sequence of curricular topics, with tests that guided and sometimes determined subsequent instruction. Propelled by behaviorist theories (Skinner, 1953), these systems included tests to check understanding and provide immediate feedback to reinforce desired behaviors.

To support these uses, tests had to support inferences about what students could do. Glaser (1963) coined the term *criterion referenced* to describe tests designed to assess

“the degree to which the student has attained criterion performance” (p. 6). This set up a useful contrast to *norm-referenced tests* that have a primary purpose of differentiating among test takers relatively, often for the purpose of selection into competitive programs. Although well-designed tests can support scores that have criterion-referenced and norm-referenced interpretations (AERA et al., 2014), essentially all modern accountability tests claim to be criterion referenced to emphasize that the primary scores of interest indicate what students know and can do against an absolute standard. The term also suggests the appealing theoretical possibility that all test takers can achieve a given criterion level.

To support or operationalize decisions in programmed instruction, practitioners and systems use established cut scores or set them on the basis of what students know and can do. Bloom’s popular “mastery learning” model (1968), for example, included end-of-unit tests with cut scores that distinguish between students who are and those who are not ready for the next unit. Cut scores on criterion-referenced tests should enable teachers to describe a student who scores at or above a cut score in terms of their absolute knowledge and skills, without reference to any norm or reference group.

Tests in this era became much more focused on fine-grained objectives that developed along continua. Score reporting turned to support interpretations of specific criteria, knowledge, skills, and abilities. Programmed instruction also raised the stakes on potential instructional decisions for students, including remediation and repetition of previous instruction. And the use of tests to reform and improve instruction, both directly by providing data and informing decisions and indirectly by focusing the system, continued to rise.

Testing for Monitoring and Program Evaluation

As programmed instruction was increasing the role of testing in classroom instruction, the 1960s saw an increase in the role of testing in federal programs and surveys. The Elementary and Secondary Education Act of 1965 (ESEA) included a required, annual, measurement-based evaluation of Title I programs under the Title I Reporting and Evaluation System. The National Science Foundation began to require the use of educational tests in evaluations of funded curricula. And the headline findings from the Equality of Educational Opportunity Survey (Coleman et al., 1966) were drawn from standardized achievement tests. Extending the Tyler rationale, educational testing began to be a default requirement for federal programs, a signal of rigor, transparency, and public accountability.

Planning for NAEP also began in the 1960s, and the first NAEP assessments were administered in 1969–1970 (Jones & Olkin, 2004). Early NAEP testing did not report scale scores and was instead inspired by the same objectives-oriented theories that guided programmed instruction. In a 2004 reflection, Cronbach compared the reporting goals to those of opinion polling, where results consisted of numerous percentages of correct answers to specific questions, disaggregated by groups. He also notes, “There was interest in making the results comprehensible to teachers in terms that they could

put to use in their teaching” (Cronbach, 2004, p. 141). However, the reports were unwieldy and made it challenging to measure progress over time. We reserve a fuller review of NAEP and its significant and growing role in educational accountability for the section “NAEP and a Theory of Public Accountability”. Here, we observe the growing pains of early NAEP as an example of a continuing tension between designing diagnostic tests for instructional guidance and designing monitoring tests for large-scale educational progress.

Minimum Competency Testing

In the 1970s, concern about youth unemployment, low graduation standards, and intractable achievement gaps helped to spur a “back to basics” movement (Haertel & Herman, 2005). In response, district and state policies known as *minimum competency testing* began to spread, typically taking the form of making high school graduation and, in some cases, grade promotion contingent on students scoring above selected score thresholds on standardized tests. Statewide minimum competency testing requirements existed in at least 29 states by the end of the decade.

Minimum competency testing built on the theoretical foundations of educational psychologists from previous decades, including the use of tests that measured specific competencies and the use of a criterion-referenced cut score for “minimum competency.” However, a dramatic extension from previous decades was the use of large-scale educational policy to articulate a clear and high-stakes decision rule, in this case with direct consequences for students. The rhetoric of the movement has largely faded in favor of an emphasis on high standards and higher-order thinking skills. Nonetheless, the movement helped to advance a more substantial state role in educational testing, as well as the use of criterion-referenced tests and cut scores for high-stakes decisions (Koretz & Hamilton, 2006).

With low cut scores and numerous opportunities to retake exams, Catterall (1989) summarized the scholarly consensus that “the competency test was legislated as a lion but implemented as a lamb” (p. 4). As we review federal and state accountability policies for schools and students over the next decades, this pattern recurs: initially unrealistic political rhetoric and goals are followed predictably by increasing flexibility.

The Birth of the Standards Movement

The 1980s began with a more centralized and data-driven effort than the previous decade to signal a crisis that required national educational reform. The influential 1983 report *A Nation at Risk* (National Commission on Excellence in Education, 1983) used a range of indicators, including international comparisons from the International Association for the Evaluation of Educational Achievement (Bloom, 1973), results from the SAT (formerly Scholastic Aptitude Test), and NAEP results to argue that “the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people” (National Commission on Excellence in Education, 1983, p. 5). The report called out the shortcomings of

minimum competency tests specifically, arguing that “the ‘minimum’ tends to become the ‘maximum’ thus lowering educational standards for all” (p. 20). The report instead called for “more rigorous and measurable standards, and higher expectations, for academic performance” (p. 27). The report specified that tests “should be administered at major transition points from one level of schooling to another” and “should include other diagnostic procedures that assist teachers and students to evaluate student progress” (p. 28).

The coming years continued to clarify the logic of modern large-scale accountability testing: aligned tests, higher standards, and transparent reporting helped to focus and improve teaching and learning (Resnick & Resnick, 1992; Smith & O’Day, 1991). The state and federal roles in accountability testing also continued to rise. Responding to requests from some governors in 1984, Secretary of Education Terrel Bell prepared a one-page “wall chart” of state educational outcomes, including graduation rates and SAT and ACT (formerly American College Testing) scores. Once planned for release in a small press conference in the secretary’s office, the wall chart resulted in the largest press conference in the Department of Education’s history (Ginsburg et al., 1988). Later that year, the Council of Chief State School Officers passed by a vote of 27 to 12 a report supporting improved measurement of educational indicators across states, with a closer 20–19 vote that avoided the deferral of a key section of the report about comparable outcomes (Toch, 1984). The report noted that, “on a technical level, state-to-state comparisons are problematical for a variety of reasons . . . on a political level, however, the attention given to the Secretary’s ‘wall chart’ makes inevitable future state to state comparisons on outcome measures” (Council of Chief State School Officers, 1984, p. 3). This articulates a theory of public accountability that we develop in the next section, where transparent reporting of comparable outcomes signals worthy goals, creates competition, and draws public attention, even without explicit consequences or rewards for scores.

In addition to higher performance standards as exemplified by “proficiency,” reports from this era emphasized complex, higher-order thinking skills and attempted to distinguish clearly between performance standards and content standards. Content standards referred to “the knowledge, skills, and other understandings that schools should teach in order for students to attain high levels of competency in challenging subject matter,” and performance standards “define various levels of competence in the challenging subject matters set out in the standards” (National Council on Education Standards and Testing, 1992, p. 13). The National Council on Education Standards and Testing was convened to explore the possibility of a national system of standards and tests, with candidate sets of content standards like those by the National Council of Teachers of Mathematics in 1989 and the American Association for the Advancement of Science in 1993.

An ambitious effort known as the New Standards Project was a joint effort of the University of Pittsburgh’s Learning Research and Development Center and the National Center on Education and the Economy. The project attempted to operationalize higher

order content standards in the form of performance assessments. The project had 17 states and 6 urban districts signed on at its peak. It ultimately faced a number of challenges, including state resistance to standards that were not “home grown” and apparent underestimation of the amount of time and money it takes to produce performance assessments that support accountability decisions at scale (Hoff, 2001). Koretz et al. (1994) summarized similar tensions among affordability, score precision, and instructional relevance in Vermont’s portfolio assessment system.

The 1990s also saw the rise of explicit school-level stakes. Whereas minimum competency testing had direct consequences on students, the logic of standards-based reform focused on incentives at the school level. The popular Quality Counts series published by *Education Week* reviewed state educational reform plans annually, including a grade for states on multiple indicators related to standards-based reform. By 1999, the report found that 36 states had committed to transparent reporting of report cards for schools, with 19 issuing overall performance-based ratings (Jerald, 2000). In addition, 16 states had the legal authority to reconstitute schools that failed performance measures, and 13 states offered monetary rewards to successful schools. These state systems established the precedent for the explicit school sanctions that would be required of all states by NCLB in the years to come.

The NCLB Era

The NCLB Act, passed in 2001, was a bipartisan triumph of the standards movement in education. NCLB represented a reaction to federal discontent over the progress of school reform throughout the 1990s, most specifically the slow and weak implementation of the prior ESEA reauthorization, the Improving America’s Schools Act (McDonnell, 2005). NCLB also grew out of state-level reforms, primarily in Texas, that seemed to be boosting student performance (Deming et al., 2016; Grissmer & Flanagan, 1998).

The NCLB suite of reforms was premised on a straightforward theory of change (Porter & Polikoff, 2007). First, establish clear, grade-level content standards in the core academic subjects (mathematics and English language arts, though science was also included in the law). Second, create aligned assessments to measure student performance against those standards and inform the public and educators of that performance. Third, through accountability policies, hold schools and districts accountable for student performance against the standards and offer interventions in schools where the standards are not met. Fourth, disaggregate all results for key student groups of interest to ensure high average performance does not mask low-performing student groups. With these policies in place, the thinking went, teachers (supported by school and district leaders) would be incentivized to align their instruction with the challenging academic content standards, student opportunity to learn would improve, and achievement would rise. In schools where achievement was not rising, public school choice (enabled by the public release of school performance data) would put additional market pressure on schools to improve, while supplemental education services would provide additional support to students and teachers. The design of the accountability

systems broadly aligned with this theory of change, though we discuss the effectiveness of these policies in the section “The Impact of Accountability Policies”.

School accountability under NCLB was based on a straightforward set of measures: student proficiency rates on state tests, student participation rates on state tests, and an additional state-selected indicator (high school graduation rates and, typically, attendance rates for elementary and middle schools). For student proficiency, states were required to establish proficiency rate trajectories for each tested subject and grade. These trajectories began from baseline proficiency rates set in 2002 based on the distribution of student performance in the state to the target of 100% proficiency by 2014. Each year’s target was called an annual measurable objective (AMO), and meeting all AMOs was necessary to make adequate yearly progress (AYP) and avoid consequences. In other words, the NCLB accountability system represented a conjunctive approach to combining multiple measures—schools had to cross each AMO bar overall, and for all student groups—high performance in one area could not offset low performance in another.

The consequences for failing to meet AYP escalated with the number of years of failure, beginning with labeling, then public school choice, then supplemental tutoring, and finally the possibility of state intervention (including school closure, charter conversion, state takeover, or other significant turnaround efforts). States were allowed to set different trajectories for different grade levels and subject areas, but they were required to have the same targets for all student groups (NCLB required states to disaggregate data for, at minimum, student groups based on race/ethnicity, socioeconomic status, disability status, and English proficiency). To ensure that schools were not selectively excluding low-performing students from taking tests, the law also required 95% of each student group to participate in each exam.

While these test-based performance measures seem straightforward and comparable across states, in practice, state decisions about the construction of their systems dramatically affected how many and what kinds of schools would be subject to accountability pressure. These decisions are described well elsewhere (Davidson et al., 2015; Polikoff & Wrabel, 2013; Porter et al., 2005), but we briefly summarize some of the relevant key decision points here.

One decision point states were allowed to make under NCLB was about student group size. This was the minimum number of students required in a particular student group for that student group to count toward accountability. The modal subgroup size under NCLB was 30 (Polikoff & Wrabel, 2013), meaning that if a school had 29 low-income students, that student group would not count as its own group for accountability purposes, whereas if it had 30 or more, it would. (The students would count in the overall school average regardless.) However, minimum group sizes ranged considerably across states (from a low of 5 to a high of 100; Davidson et al., 2015). Smaller student group sizes can dramatically affect the number of students who are included in student group accountability measures. An analysis of California data (Polikoff et al., 2018), for instance, found that moving from their old group size of 100 to a group size of 30

increased the average number of accountable student groups in schools from 3.65 to 5.44. This had the effect of dramatically reducing the number of students who were in groups not reported or used for accountability (for instance, under the previous system, 80% of multiple-race students were in schools where that group was not large enough to be reported, but under the new system it was just 28%). Of course, there are trade-offs of changes in student group size rules: With smaller minimum group sizes, reliability will be lower (year-to-year fluctuations will be higher), but there will be greater incentive for educators to attend to student group performance if the school is accountable for more student groups.

Second, as described by Polikoff and Wrabel (2013), the law allowed states to use a variety of alternative methods in determining whether schools' and districts' proficiency rates were above the performance target. By far the best known of these was the Safe Harbor provision, where schools were given credit for meeting a proficiency target if the percentage of students scoring below *Proficient* decreased by 10% (not 10 percentage points) from the previous year. These alternative methods had the effect of sometimes dramatically lowering the proficiency threshold. For instance, consider a school with 100 students, of whom 50 are *Proficient* (50% proficiency rate) against an AYP target of 75%. This school would need to get just 5 more students over the proficiency threshold in the next year to meet the AYP target, even though the proficiency rate of 55% would fall far below the actual target for that year.

Third, as described in detail by Davidson et al. (2015), states also could use statistical adjustments to their accountability results in making accountability decisions. Specifically, states were permitted to adjust targets based on confidence intervals, in essence allowing states to say, "we will only hold your school accountable if we are extremely confident your school's performance is below the proficiency threshold." Confidence intervals were defined in terms of sampling variance of proportions, but were applied in an unusual way. Instead of indicating the precision of the parameter estimate, they were used to lower targets in proportion to imprecision. Importantly, confidence intervals and Safe Harbor could be layered on top of one another. In the previous example of the 100-student school in California, it would actually only need to get 2 more students over the proficiency line, because California applied a 75% confidence interval to the 5-student Safe Harbor target.

Fourth, states had leeway over a variety of technical decisions about which students would count for accountability and how to aggregate their scores. For example, states applied variable rules around the inclusion of students with certain kinds of disabilities and varied in their selection of student racial/ethnic groups (Davidson et al., 2015). States could set their own rules about student enrollment, with some states requiring students to be enrolled in a school for an entire year to count for accountability purposes and others requiring just a few months' enrollment (the former approach excluding highly mobile students). And states could decide how to aggregate results across grades and subjects, whether by first aggregating results and then comparing these to AYP targets or by comparing results to targets separately by grade and subject.

Finally, states were allowed to set nonlinear performance trajectories from the 2002 starting point to the 2014 100% proficiency target (Porter et al., 2005). Some states set “straight-line” targets, expecting a constant annual increase in the percentage of students scoring *Proficient*. Other states set more back-loaded targets, with smaller annual increases in the earlier years of the law and larger increases in the later years.

These state policy decisions—some of them seemingly modest—had profound effects on the number and type of schools identified as failing to meet NCLB proficiency targets. For instance:

- Porter et al. (2005) found early in NCLB that Kentucky’s design decisions (using a large student group size of 60 and confidence intervals) resulted in 75% fewer schools failing to meet their accountability targets than had the state made more stringent decisions.
- Polikoff and Wrabel (2013) found that two thirds of California schools used Safe Harbor to meet at least one AYP AMO in 2011.
- Davidson et al. (2015) documented the complex and interacting effects of state accountability policy decisions on failure rates, suggesting that school failure under NCLB was as much about minor differences in the ways state policy makers interpreted and implemented the law as it was about actual school performance.

As the NCLB era drew toward its conclusion, there was widespread dissatisfaction with NCLB accountability systems, with specific focus on (a) the unreasonableness of 100% proficiency targets (though again, in practice there were many ways to make AYP without hitting 100% proficiency), (b) the use of status measures of performance without considering growth, (c) the narrowing effects of the law resulting from its near-exclusive focus on English language arts (ELA) and mathematics test scores, and (d) the perceived ineffectiveness of NCLB’s mandated interventions.

The ESEA Waivers

It was in this context that in 2011 (already 4 years after the original intended ESEA reauthorization), the U.S. Department of Education began offering states flexibility waivers from some provisions of NCLB’s accountability rules. The waiver accountability policies were broadly aligned with NCLB policies and the NCLB theory of change, but offered some technical fixes aimed to address the shortcomings of NCLB accountability systems.

In place of NCLB’s AYP rules, the waivers had several requirements for state accountability systems. First, state AMO targets had to be revised, with the waivers offering states two options: AMOs that increase in equal annual increments to 100% proficiency by 2019–2020 or reduce by half the percentage of *Below Proficient* students in the “all students” group and in each student group within 6 years. These tweaks were offered to reduce the pressure of the impending 100% proficiency deadline, which was then just a couple years off. Second, states were required to include a measure of student growth

in their accountability rules to address the widely known concerns about student proficiency rates as a measure of school performance. Third, states were mandated to identify at least 15% of schools for intervention, with 10% of schools identified as “focus” schools (intended to be schools contributing to state achievement gaps) and 5% of schools identified as “priority” schools (intended to be the state’s lowest-performing schools). This provision offered states some relief in terms of the number of schools and districts they would have to identify and work with, but it also introduced a norm-referenced approach into what was originally intended to be a criterion-referenced policy system based on schools meeting targets. Though the waiver accountability systems were in place for only a few years, there are generalizable lessons in the revisions about how to set realistic targets and establish consistent definitions to avoid perverse incentives.

Waiver accountability systems represented important improvements over NCLB systems in some ways, but less so in others. Polikoff et al. (2014) reviewed these improvements and missed opportunities, which we summarize here. First, waiver accountability systems offered flexibility to include measures beyond test scores, responding to earlier critiques of NCLB as narrowly focusing on ELA and mathematics proficiency at the expense of all other measures of school effectiveness. States used this flexibility to include a variety of measures like graduation rates, attendance, school climate, and opportunity-to-learn measures. Only five states used just test scores in identification of their priority schools, and most states used at least one nontest measure in their identification of focus schools. Waiver plans also often included additional tested subjects in accountability (e.g., science, which was required to be tested but was not used for accountability under NCLB), further broadening the measures against which schools were evaluated.

Second, when waiver accountability systems did rely on test scores, they included measures that experts argued were superior to proficiency rates in terms of measuring and incentivizing school effectiveness (e.g., Ho, 2008; Linn, 2003; Polikoff & Wrabel, 2013; Ryan, 2004). About half of states used a measure of individual student achievement growth, and about a quarter used different status-based measures of performance, such as performance indices where students scoring *Basic* earned half a point and those scoring *Proficient* earned a full point (Polikoff et al., 2014). Third, waiver plans made more stable classifications—identifying schools and working with them for at least a few years before moving on to a new set of schools. This has two potential benefits: It reduces the extent to which ephemeral fluctuations in performance drive school classifications and it ensures more sustained intervention and support in struggling schools.

However, researchers also raised concerns about the design of waiver accountability systems. These concerns were along a number of dimensions. For example, in some states the state would use one approach to rate schools on, say, an A–F composite index, but then use a different approach altogether to identify schools for intervention. This created a lack of consistency that might cause unnecessary confusion. As another example, while nontest measures were included in many states’ accountability systems,

proficiency rates based on state tests still dominated accountability classification rules in nearly all states, raising all of the same issues as discussed in the NCLB era. And finally, some were concerned about the use of so-called super-subgroups, which combined the performance of multiple groups of students, rather than fully disaggregating results as under NCLB (e.g., R. A. Hernández, 2013). There was concern that this approach to aggregating student groups could mask the performance of the disaggregated groups and therefore diminish attention to supporting these student groups.

Overall, the waiver period represented a controversial but important transition from NCLB-style accountability to ESSA accountability. The waivers clearly broadened the measures against which schools were evaluated, a transition from the nearly exclusive test-based focus of NCLB. The waivers narrowed attention to a small proportion of the lowest-performing schools, moving away from NCLB's extraordinarily high school failure rates and toward a more norm-referenced approach to school accountability. The waivers encouraged innovation in the use of student achievement test scores, including a move away from proficiency-based status measures and toward true student growth measures. These innovations carried over when ESEA was finally reauthorized as the ESSA in 2015.

The Every Student Succeeds Act

The ESSA passed in 2015, finally reauthorizing ESEA and updating requirements and expectations with regard to state accountability systems. The act's accountability provisions continued and expanded on some of the main innovations of the waiver era and continued the rollback of NCLB-era sanctions. The proportion of schools subject to accountability sanctions was reduced to just the bottom 5% (plus high schools with low graduation rates), and the sanctions were made much less prescriptive (with a focus instead on district creation and implementation of evidence-based plans for school improvement).

The statute describes five types of indicators that must be included in school accountability systems (Marion, 2016): first, an indicator of academic achievement; this can be as simple as percent *Proficient*, but ESSA also allows for states to use other status measures of performance (e.g., average scale scores, index-based approaches); second, another “valid and reliable academic measure,”¹ often a measure of either growth or achievement gaps; third, graduation rates, including extended (i.e., 5- or 6-year) graduation rates; fourth, progress toward English language proficiency for English learners; and fifth, another indicator of school quality or success that meaningfully differentiates and is valid, reliable, and comparable (in the text of the law, they list as options measures of student engagement, educator engagement, student access to and completion of advanced coursework, postsecondary readiness, and school climate and safety). In terms of actual accountability classifications, the statute also says that states must measure school performance against state-determined status and improvement goals in at least the areas of academic achievement, graduation rate, and student groups that are behind (Marion, 2016). Drawing on the waivers, the law requires that states produce

at least every 3 years a list of schools for comprehensive support and improvement, including (a) the lowest 5% of Title I schools, (b) high schools with graduation rates below 67%, and (c) schools with low-performing student groups (thus including both norm- and criterion-referenced approaches to school identification).

How did states implement these requirements? The approved state ESSA plans varied along every dimension, so characterizing them concisely is difficult. Indeed, if there is one way to characterize state ESSA accountability systems, it is that they are incredibly variable in form and content. But a few trends emerged in the many external reviews that were produced.

With regard to student achievement measures, many (but not a majority of) states use status measures of student performance that move beyond percent *Proficient* (e.g., average scale scores or “indices” that use a fuller range of student scores rather than a binary *Proficient/Nonproficient* distinction) and many also include achievement in subjects other than mathematics and ELA as another indicator of school quality or student success (English, 2017). Virtually all states (47 + Washington, DC; Education Commission of the States, 2018) include measures of student achievement growth in their ESSA systems, a sharp departure from waiver policies (American Institutes for Research, 2017; English, 2017). States also include test participation in their achievement measures, most commonly by assigning a score of zero to nonparticipants on state tests (English, 2017). The story for achievement measures at the high school level is even more variable from state to state; see Petrilli et al. (2016a, 2016b).

State goals and approaches also vary considerably with regard to graduation rate and other high school measures. State 4-year graduation rate goals range from a high of 100% (South Dakota) to a low of 84% (Nevada, Virginia), with the most common levels being 90% or 95% (21 and 10 states, respectively; Hackmann et al., 2019). States also include a wide variety of “college- and career-readiness” (CCR) measures, mostly at the high school level, but sometimes at lower levels. The most common of these are Advanced Placement/ACT/SAT/International Baccalaureate exam scores, industry-recognized credentials, Career and Technical Education (CTE) course sequence completion, advanced diploma completion, and other measures of course taking (for a full list, see Hackmann et al., 2019). In most but not all states, these measures are reported overall and separately for student groups. In many states, these diverse measures are aggregated into a single index (e.g., in Arizona, students can earn fractions of a CCR point for various accomplishments, with those student scores averaged across the state). In most states, the aggregated CCR measure is then combined with other measures to create an overall summative rating.

Another important trend is with regard to the “other academic indicator.” While 33 states include the CCR measure as part of their other indicators (Education Commission of the States, 2018), many alternative accountability measures are also included, representing a continued growth in the number and variety of nontest measures that began under the waivers. For instance, the majority of states include chronic absenteeism as a measure (English, 2017), which is important because absenteeism is highly predictive of other long-term student outcomes (Ansari & Pianta, 2019; Liu et al.,

2021). Other measures in this area include conditions of learning or school climate, some measure of the well roundedness of students' education, and credit accumulation.

Student groups continue to play an important role in accountability systems, but perhaps less so than under NCLB. Although ESSA requires states to continue to report student group performance, it does not require a particular approach to the use of student group performance for accountability purposes. Some states under ESSA are using a variant on super-subgroups that began under the waivers, most often simply a group defined by the lowest-performing students (English, 2017). But other states combine low-performing student groups into a demographically defined student subgroup. Furthermore, state minimum subgroup size rules continue to decrease under ESSA, with a median and modal value of 20 (Alliance for Excellent Education, 2018).

Finally, a large majority of states combine multiple measures in some way to produce a summative rating of some kind. Thirty-six states use an A–F grade, a descriptive rating system (e.g., *Needs Improvement, Average, Good, Great, Excellent*), or a numerical index system (e.g., 1–10, 1–100; Education Commission of the States, 2018). These indices vary along every conceivable dimension in terms of how they aggregate and weight their various component measures.

While most reviews of ESSA accountability plans view those plans positively, there are some concerns that are commonly raised. The most common critiques have to do with the treatment of student groups. While NCLB was critiqued by some for its overly harsh attention to student groups (its conjunctive approach meant that a single group failing a single AMO would lead to the whole school failing), ESSA may have swung the pendulum far in the opposite direction. Independent reviews suggest that most state plans give almost no attention to the performance of student groups, with school ratings being determined on whole-school performance (Bellwether Education Partners, 2017). Advocates for students with disabilities have expressed similar concerns, worrying that these students will fall through the cracks of the 33 state accountability systems that do not explicitly take student group performance into account in rating schools (National Center for Learning Disabilities, 2018). Another critique is that states' composite indices may appear transparent on the surface, but actually are overly complicated in ways that make their results confusing to educators and the general public (Bellwether Education Partners, 2017). There are also continued concerns about high school accountability, which is heavily reliant on graduation rates (a status-based measure of performance) and lacks the growth-based approach used when student-level longitudinal test score data are available. It is also not fully clear in many state plans how states actually plan to intervene in and improve their lowest-performing schools (though the literature does not exactly offer a road map for school improvement at scale).

The overall takeaway from ESSA plans can be summarized as follows. These plans expand on some of the trends that began under the waivers. In most cases, they dramatically increase the types and number of measures to be used in accountability, but proficiency rates in mathematics and ELA still play a major role and may drive educators' actions. They more appropriately measure student achievement status and

almost universally include measures of student achievement growth as well. However, they take some of the pressure off for school improvement (which is by design and a reaction to NCLB), substantially decreasing the number of schools that are likely to feel meaningful accountability pressure. The changes to student group reporting and accountability requirements may reduce awareness about the average achievement and progress of historically underserved students when these populations are small.

There are two other ESSA-era trends that are worth mentioning here. The first concerns changes in the high school summative assessment. As of 2019, half of the states were including either or both the SAT and the ACT as a high school summative measure (Olson, 2019). This change was allowed under the law's provisions for "a nationally recognized college entrance exam," and states were enthusiastic about the opportunity to adopt assessments that could serve multiple purposes. There are continued concerns about the alignment of the tests to state standards (e.g., National Council on Measurement in Education, 2019), but the Department of Education approved the use of the tests for statewide summative purposes regardless.

Second, as ESSA accountability systems were being implemented, the COVID-19 pandemic gripped America's schools in the 2019–2020 academic year. There were important implications of the pandemic on assessment and accountability. For instance, the Department of Education authorized states to extend the testing window, shorten assessments, and give assessments remotely (U.S. Department of Education, 2021). The Department of Education also allowed states to waive the 95% participation rate rules and to apply for waivers to the requirements for school identification. The Department of Education has been returning to ESSA accountability rules as the pandemic abates.

NAEP AND A THEORY OF PUBLIC ACCOUNTABILITY

Public accountability includes policies and commitments for reporting test data without predetermined stakes attached to results (Hutt & Polikoff, 2020). As Figure 16.1 indicates, proponents of public accountability assert or assume that publishing test score data can signal worthy goals, improve commitment to those goals, spur competition toward those goals, encourage actors to learn from organizations that are successful at achieving those goals, and align the system toward those goals. This logically requires appropriate measures that communicate effectively to intended audiences. The reasoning also requires that actors are motivated by public transparency without any automatic sanctions or rewards attached to results.

This section reviews the rise and species of "report cards" that represent the linchpin of test-based public accountability. We begin with the assessment known as the Nation's Report Card, NAEP (<https://www.nationsreportcard.gov/>), which was introduced briefly in the previous section. NAEP remains the foremost national and cross-state example of test-based public accountability. This section continues with a discussion of audiences and effects of public reporting.

The Rise of NAEP as a Mechanism for Public Accountability

NAEP had its origins in the federal Office of Education under the leadership of Francis Keppel in 1969. The earliest NAEP assessments that year tested citizenship, science, and writing skills of 9- 13-, and 17-year-olds, debuting technical features like matrix sampling. Over subsequent years, new innovations were included, such as comprehensive content frameworks, state comparisons, and, later, large urban district comparisons. In 1986, Secretary Bennett formed a study group chaired by Tennessee governor Lamar Alexander and Spencer Foundation president H. Thomas James to explore state comparisons using NAEP (Alexander et al., 1987). The resulting report compared NAEP to a weather map that only provided weather for the nation as a whole. The report described state-to-state comparisons as “the single most important change recommended by the Study Group” (p. 11). This recommendation was fiercely debated and ultimately included in the 1988 amendments of ESEA as a trial state assessment program beginning in 1990 that would be voluntary for states (Vinovskis, 2001). Later, NCLB required states to participate in NAEP biennially as a condition of receiving Title I funds.

NAEP also illustrated the value of compelling reporting metrics for public accountability by adopting the criterion-referenced score metric, “percent *Proficient*.” The 1987 Alexander–James report included a published commentary by a National Academy of Education review committee chaired by Robert Glaser, the same scholar who coined the term *criterion-referenced testing* 24 years prior. Although the Alexander–James report did not mention criterion-referenced score reporting, the commentary does:

We recommend that, to the maximal extent technically feasible, NAEP use descriptive classifications as its principal reporting scheme in future assessments. For each content area NAEP should articulate clear descriptions of performance levels, descriptions that might be analogous to such craft rankings as novice, journeyman, highly competent, and expert. Descriptions of this kind would be extremely useful to educators, parents, legislators, and an informed public. (Alexander et al., 1987, p. 58)

The subsequent 1988 legislation created the independent NAGB to oversee NAEP policy and assigned to NAGB the responsibility of “identifying appropriate achievement goals for each age and grade in each subject area to be tested” (PL 100-297, Part C, Section 3403(6)A, 1988). Shortly thereafter, in 1989, President George H. W. Bush convened all 50 governors in an education summit to achieve consensus on broad educational goals for the year 2000. In 1989, the National Governors Association established Goal 3: “American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter” (National Governors Association, 1990). A year later, in May 1990, the newly formed NAGB voted to establish three achievement levels for each grade and subject, with a central level of *Proficient* including the same descriptor: “Students reaching this level have demonstrated competency over challenging subject matter.” Although the NAGB process for setting performance

levels for “proficiency” faced considerable scrutiny and controversy through the 1990s, (Shepard et al., 1993; Vinovskis, 1998), a 2017 evaluation concluded, “Through 25 years of use, the NAEP achievement levels have acquired a ‘use validity’ or reasonableness by virtue of familiarity” (NASEM, 2017, p. 11). The apparent interpretability of the 0- to 100-scaled percent *Proficient* statistic would help it to become a central feature of test-based public accountability policies in the decades to come, despite its technical flaws (Ho, 2008; Polikoff, 2016).

Although official reporting of NAEP scores includes states and participating large urban districts (from 6 in 2002 to 27 in 2019), NAEP data form the basis of a number of aggregate mapping efforts that represent a form of public accountability. The National Center for Education Statistics has regularly published a mapping of state performance standards to the NAEP scale (e.g., Bandeira de Mello et al., 2009; Ji et al., 2021), revealing wide variation in the rigor of proficiency standards across states under NCLB. These mappings show the rise and convergence of proficiency standards through the ESSA era. NAEP scores also form the basis of linking efforts that attempt to map school and district test scores to a common national scale for aggregate research and public engagement (Reardon et al., 2021). These secondary research efforts extend the reach of test-based public accountability even as they lag years behind the time of testing.

NCLB and ESSA required states to participate in only two NAEP subjects, reading and mathematics. Additional NAEP subjects have included science, economics, arts, foreign language, civics, geography, and technology and engineering literacy, all typically reported at the national level only. NAGB interest and budgetary constraints have led to less frequent testing in these subjects in recent years. This is consistent with a theory of public accountability: Absent required participation, competition, and alignment with state goals, tests at other levels of the system will be less effective at inspiring progress.

Hutt and Polikoff (2020) distinguished between a range of variables for public accountability, not only test-based variables but also curriculum materials, budget, and salary data. They classified NAEP as relatively inactionable compared to other variables because of NAEP’s high level of aggregation and delayed reporting schedule. In the remaining subsections, we review other audiences for public accountability that might use report cards like NAEP but at different levels of aggregation and for different purposes.

Forms of Report Card Accountability

There is a long history of states implementing report card accountability, where they publish publicly available report cards of school performance indicators. These report cards are intended to provide information to parents and stakeholders to both inform parent choices and encourage external pressure for schools to improve and meet benchmarks. These “public accountability” systems began to be implemented in the mid-1990s (Hanushek & Raymond, 2005), such that by the onset of NCLB in 2002, about

20 states had report card–style accountability systems without attaching consequences. (Of note, early research on report card–style accountability systems found they did not have the same positive effects that consequential systems did; Hanushek & Raymond, 2005). Of course, states with consequential accountability systems might also put out report cards, and indeed, this became required under NCLB.

Under ESSA, all states have accountability report cards that summarize school and district performance and compare them to state benchmarks. State (and Washington, DC) report cards and dashboards vary tremendously in the ways in which they present accountability data. Of the 51 report cards produced, 26 have some form of summative rating. Of these, 12 give schools a letter grade, 8 a numerical score, 7 a categorical label (e.g., “*Commendable*”), and 5 a star rating (some states give schools multiple summative ratings). The Data Quality Campaign (2019) offers a high-level overview of progress on state accountability report cards each year. In their report, they highlighted strengths and weaknesses of report cards across the nation. Strengths include that the majority of state report cards are easily found through Internet searches, are in formats that are useful to parents (PDF and mobile friendly), and include downloadable data. Weaknesses include that most state report cards do not offer translation to any language other than English, include text at too high a reading level, lack disaggregated data for at least one federally required student group, and lack reporting on important measures like college enrollment and teacher demographics and effectiveness. In addition, the Education Commission of the States (2014) offered a more detailed summary of the similarities and differences in report cards as of that time.

Another form of reporting educational data is through public-facing “dashboards,” which display school and district performance along multiple dimensions. California adopted a school dashboard, which they have modeled on dashboards from other jurisdictions (Darling-Hammond et al., 2014). In California’s dashboard (Polikoff et al., 2018), schools are given ratings for “status” and “change” on each of several dimensions (including ELA and math achievement, chronic absenteeism, suspension rates, and English learner progress) on a five-point color-coding rating scheme. These public-facing dashboards are intended to drive parent decisions and involvement in states’ (nonconsequential) public accountability system, the Local Control and Accountability Plans, but evidence suggests that few access these data (Polikoff et al., 2018) or participate in these processes (Marsh & Koppich, 2018).

How Parents Use Accountability Data

One of the main targets of accountability data is parents. Providing accountability data to parents could affect performance in several ways. For an individual parent, receiving accountability data and using it to select more effective schools for their child could result in improvements in that child’s performance. Accountability data use by parents could also affect the broader system. Or if enough parents made school choices based on accountability data, they could drive systemic improvements in school

performance by leading to the growth of more effective schooling options or the closure of less effective ones (for more on this topic, see the literature on “exit” and “voice,” for instance, Hirschman, 1970; for a recent critique of the theory by which public accountability of this form can result in educational improvement, see Hutt & Polikoff, 2020). And, of course, individual parents might also use their child’s state test results to press for educational improvement for their individual child (e.g., needed instructional supports, acceleration opportunities).

School accountability report cards themselves have not been the subject of much empirical study, but there are a few exceptions. A study by the Institute of Education Science’s National Center for Education Evaluation and Regional Assistance (2019) tested the impact of five different features (e.g., the balance of numbers and graphs, the ordering of schools by distance or performance, and whether district averages were included) on school report cards on parents’ knowledge, satisfaction, and school choices in an experimental simulation study. They found evidence that these features matter (though to a fairly modest extent), with each feature mattering in different ways. For instance, they found that presenting accountability results as numbers only increased parents’ ability to accurately answer factual questions about school performance indicated on the report card, but presenting numbers and graphs increased parent satisfaction. They also found that including parent satisfaction in accountability report cards boosted parents’ satisfaction, but slightly diminished their knowledge. Finally, they found that adding more information to the presented results boosted parents’ satisfaction but may have led them to select fewer effective schools. The Education Commission of the States (2014) also analyzed parents’ preferences with regard to school report cards and accountability data, finding that ease of interpretation, thoroughness of data (including nontest measures), and ability to compare schools were especially important in parents’ evaluations of report card quality. Finally, Jacobsen and colleagues (2014) tested certain features of accountability report cards more formally, finding that the form by which results are reported (e.g., A–F, proficiency rate) substantially affects voters’ satisfaction with high- and low-performing schools (for instance, letter grades widened the perceived quality gap between high- and low-performing schools as compared to proficiency rates).

There is also more general literature on how parents evaluate schools. This literature is not necessarily about accountability reporting per se, but their results speak to the kinds of measures that might be valued by parents on accountability report cards. Haderlein (2022) divided these literatures into stated preferences and revealed preferences. Stated preferences research relies on parent self-reports about the factors they think are most important in evaluating schools. Revealed preferences research estimates parent preferences without explicitly asking them (i.e., by analyzing their behaviors).

Survey studies of parents suggest that they value academic quality measures the most when evaluating schools. The particular academic quality variables range by study, but include teacher quality (Schneider et al., 1998), curriculum quality (Van Dunk

et al., 1998), and academic outcomes such as test scores (Armor & Peiser, 1998). Some studies suggest that there are heterogeneous preferences for these factors (with low-income and less-educated parents expressing even stronger emphasis on academic measures; Schneider et al., 1998). There is also some evidence that school demographics may play a role in parents' decisions, especially for White parents (Holme, 2002; Roda & Wells, 2013).

Revealed preferences research generally relies on housing data, school choice preference data, or experimental data (Haderlein, 2022). Housing studies confirm that parents are willing to pay more for access to schools that have higher average academic achievement (Bayer et al., 2007; Black, 1999; Figlio & Lucas, 2004). Further, Figlio and Lucas (2004) found that accountability report card results affect housing values, with parents willing to spend more money to live on the boundary of an "A" school versus a "B" school. In the school choice context, studies using parent choice data reveal that parents value a variety of factors, including average student achievement, growth in student achievement, proximity to home, racial demographics (preferring schools with demographics closer to their own), and other nonacademic factors like extracurricular activities (Glazerman & Dotter, 2017; Harris & Larsen, 2017; Hastings et al., 2009). Recent evidence suggests that parents' perceptions of school effectiveness may actually be proxied by their perception of students' peer quality (Abdulkadiroglu et al., 2017). Finally, experimental survey studies show that parents value demographics when making school decisions, but that academic achievement levels and growth are the two most important factors (Billingham & Hunt, 2016; Haderlein, 2022). A recent and innovative study by Haderlein (2022) tested parents' choices in three ways, finding that the ways they evaluate school quality do not always line up with how they make choices among schools. In particular, parents value academic achievement growth the most when choosing among schools even though the factors contribute equally to their rating of school quality. Nontest measures also matter in parents' decisions, though more so in their rating of schools than in their selection of schools.

Taken together, the existing literature clearly demonstrates that school accountability data are important for affecting parents' evaluations of school quality. Parents clearly seem to value student performance data, including both achievement status and growth. They also seem to value nontest measures to the extent they are available, especially information on teacher quality and instructional programs. In addition, parents seem to respond to student demographics by choosing schools that are more racially similar to their own race/ethnicity. While there is a very small literature on the ways in which accountability data are presented and the impacts of that presentation on preferences and decisions, that literature does suggest that key features of data presentation matter for parents' judgments, perhaps especially at the tails of the performance distribution. Unfortunately, research also suggests that more advantaged parents are more likely to access and use accountability data, implying that more needs to be done if the data are to be useful to diverse audiences.

How Teachers Use Accountability Data

Accountability data could also be used by teachers to improve instructional quality and, through it, student performance. For example, receiving data on student performance could help teachers identify gaps in performance that need to be addressed (e.g., topics or skills where students are performing poorly or well). These data could also be used to group students (heterogeneously or homogeneously) or to target interventions. There is a growing body of literature on “data-driven decision-making”—the ways that educators use data for instructional improvement (see Marsh et al., 2006, for an early review or Marsh & Farrell, 2015, for a more recent summary). This section does not review this entire body of literature, but rather summarizes some of the key findings as they pertain to the role of state accountability performance data in teachers’ decision-making.

There is broad agreement that state-level achievement data often arrive too late to be instructionally useful (see Marsh et al., 2006). This has led to a proliferation of district-level assessment systems that are intended to provide more actionable evidence about student performance on a timeline that is more usable by educators; these are very widely adopted now, especially in urban districts (Burch, 2010). Indeed, the creators of summative state tests have recognized the importance of (and market for) aligned interim assessments. This has led to, for example, the Smarter Balanced Assessment Consortium creating and offering interim assessments that are more seamlessly integrated with summative state tests.

More generally, researchers have identified the key factors driving the use of educational data by educators. As just mentioned, one dimension is timeliness—if the data do not arrive sufficiently quickly, they cannot inform instructional decisions. Other key dimensions include the following (Marsh et al., 2006):

- accessibility of data, including technological capacity;
- quality of data, including both the reality and the perception of data quality (i.e., whether results accurately reflect performance);
- motivation to use data, including both internal and external forms of motivation;
- educator capacity to understand and use data, including the degree to which educators have been sufficiently supported on data use;
- educators’ other beliefs and values, including their beliefs about the abilities of their students (Bertrand & Marsh, 2015);
- organizational culture and leadership, including school norms, presence of collaborative opportunities, and leadership modeling effective data use;
- organizational barriers, including pressures from the curriculum and adequate time to interpret and use data; and
- local and state context, including the history of data use and accountability.

State summative accountability results may be less useful for educators than more locally developed performance data because of their timeliness and their perceived distance from instruction.

The use of performance and accountability data by teachers can also have unintended consequences (these are discussed below in the section on accountability effects).

Whether in the context of state summative assessment data or interim/benchmark data, the results of these assessments can sometimes be used to “game” accountability systems by targeting students who are close to accountability thresholds. These so-called bubble kids are the ones whose performance is most likely to drive accountability ratings, so it is a rational (if potentially pernicious) response for educators to attend to them. There is some evidence that these responses are widespread (Booher-Jennings, 2005; Marsh et al., 2006), though the actual impact on student performance of these kinds of approaches is unclear. We discuss other unintended instructional responses in the section “The Impact of Accountability Policies”.

How Researchers and Policy Makers Use Assessment Data From Accountability Systems

Assessment data from state accountability systems offer comparable performance metrics across school systems and over time. This comparability enables researchers to link relative improvements in test scores to resource allocations and policies. Phillips et al. (2018) laid out a typology of different uses of assessment data for research and policy. One type of use is to inform service provision, including to help coordinate services across government agencies (e.g., schools, social work). Another type of use is to generate descriptive evidence to support continuous improvement efforts (for example, see the University of Chicago Consortium on Schools research that has informed Chicago Public Schools’ policies and practices for several decades, e.g., Allensworth & Easton, 2005; Roderick et al., 2014). A third type of use is to identify schools or districts that are performing particularly well or poorly and use them to inform improvement efforts more broadly (e.g., Cannata et al., 2017). Fourth and finally, state longitudinal data systems have been used to conduct large amounts of quantitative research evaluating specific policies and programs, ranging from teacher evaluation to accountability and curriculum reforms.

Several states have been national leaders in the creation and dissemination of educational data for research, and as a result these states are overrepresented in the literature using statewide longitudinal data systems to evaluate policy. Phillips et al. (2018) summarized these states, which include Florida, North Carolina, Texas, and Washington. State creation and maintenance of longitudinal data systems grew sharply under the Obama administration, which used Race to the Top and substantial federal dollars to incentivize states to beef up their data systems (Howell, 2015).

Another potential use of accountability data by policy makers is for continuous improvement; this is a focus of policies like California’s Local Control and Accountability Plans. Hough et al. (2018) identified three key issues in the use of educational data for improvement processes. First, data use for improvement is different from data use for accountability purposes. In particular, educational data must be a part of a change-oriented pathway, which involves school and district leaders evaluating changes, running predictive analytics, and establishing priorities (Yeager et al., 2013). Second, data use for improvement must be embedded in an aligned improvement process. This includes

attention to establishing cycles of inquiry; building norms, culture, and mindset; and fostering organizational capacity (Marsh et al., 2006). Third, the specific data needed for continuous improvement vary by user and phase of the improvement process. State accountability data are not nearly enough to drive improvement in this context—more and different kinds of data are needed, and they must be provided in usable forms (The Victorian Quality Council, 2008).

In general, for data to be more useful for research and policy, Phillips et al. (2018) suggested that several factors are important. First, data are more useful to the extent that they are well linked over time and across agencies (e.g., across agencies at a given age level and longitudinally from early childhood through higher education and beyond). Second, data are more useful to the extent that they are clean and straightforward to use. Third, data are more useful to the extent that political barriers do not get in the way of their use. For instance, laws that prohibit the connection of student data with individual teacher identifiers (as in California) will limit the ability to conduct analyses related to teacher policies. Fourth, data are more useful if there are not too many administrative and bureaucratic requirements that limit the ability of researchers to publish their results (e.g., onerous data use agreements).

STUDENT AND TEACHER ACCOUNTABILITY IN K–12 EDUCATION

At lower levels of the system in Figure 16.1 are student and teacher accountability systems. These systems contrast with school accountability policies and public accountability strategies by using student test scores to inform decisions with direct impact on individual students and teachers. The intended direct mechanisms include matching students and teachers with curricular or professional experiences appropriate for their current skills (e.g., sorting students or teachers, assigning them to programs, awarding them credentials). Their intended indirect mechanisms include directing student and teacher effort toward desired goals, as well as indicating who holds responsibility, management, and control over key outcomes (Airasian, 1987).

The Main Types of Student Accountability Policies

Broadly, there are two main types of high-stakes student accountability assessments in use across the United States in the early 21st century: high school exit exams (HSEEs), which are sometimes delivered as comprehensive exams covering multiple subjects and sometimes delivered as end-of course (EOC) exams, and grade promotion exams.

High School Exit Exams

HSEEs are a policy by which students must pass one or more standardized tests to receive a high school diploma. Exit exams have evolved over time from the minimum competency testing (MCT) movement of the 1970s (Chudowsky et al., 2002). Like those tests, cut scores were set at a low level of difficulty and were more often used

to identify and remediate low-performing students (Dee, 2003; Jacob, 2001). As the standards movement grew in the 1990s and into the 2000s, more states began implementing HSEEs, and those with MCTs began increasing the rigor of these assessments. By the resurgence of the movement around 2010, the majority of states had implemented high-stakes HSEEs (Center on Education Policy, 2012). These exams were especially common in states serving larger proportions of non-White students—in 2011–2012, for instance, 85% of Hispanic students attended school in states requiring HSEEs, as compared to 69% of students overall. Since that time, the number of states requiring an exam to graduate has declined—at last count for 2018–2019, just 13 states still had a traditional exit exam in place (Gewertz, 2020).

Even as cut scores have risen above historical levels, they remain at relatively low levels of difficulty, often representing middle school or early high school content (Achieve, 2004). Furthermore, the vast majority of exit exam policies have come with numerous escape valves or alternative methods to pass the assessments. All states that had exit exams offered students the opportunity to retake the exams if they failed, with states allowing anywhere from 2 to 12 retakes (Center on Education Policy, 2009, 2012). More than 80% of states also offered alternative pathways to meet the standards—for instance, many states allowed test scores from other tests such as SAT or ACT to count, permitted students to use portfolios or projects to substitute for test scores, or offered waivers or appeals for students who met other requirements. Finally, nearly all states specifically offered alternative pathways for students with disabilities (Center on Education Policy, 2009, 2012). Given these alternative pathways, the actual impact of the test itself on students' graduation may have been moderated somewhat (Harris et al., 2020).

Beginning late in the NCLB era and continuing to the present, states began to move toward EOC exams to replace more traditional exit exams, recognizing that student pathways are increasingly differentiated in high school. EOC exams differ from traditional exit exams (which are sometimes referred to as “comprehensive” exit exams) in two key ways (Center on Education Policy, 2008; see also Domaleski, 2011, for a comprehensive discussion of EOC exams). First, EOC exams are intended to assess student mastery over specific course content, as opposed to providing a more general assessment of student knowledge across subjects. Second, EOC exams are given to students when they complete particular courses, rather than at a fixed point (i.e., most exit exams are given starting in 10th grade).

An analysis by Tyner and Larsen (2019) documented the rise of EOC exams across the United States. In 1996, just 2 states had EOC exams. Over the next 12 years, the count rose an average of 1 state per year to 14 states by 2008. The next 9 years saw more rapid spread, with 30 states having adopted one or more EOC exam by 2017. Since then, the count has ticked down to 26 states (not counting any COVID-related pauses). Importantly, EOC exams are widely used for both student and school accountability, and their school accountability uses (i.e., as part of high school accountability metrics) outnumber their student accountability uses.

Early advocates for EOC exams tended to emphasize that they were more appropriate than HSEEs as tools for accountability because they were better aligned with the high school curriculum (Center on Education Policy, 2008). Furthermore, they offered the opportunity to provide feedback to teachers and students about student performance to drive targeted supports and remediation activities to students. Indeed, in some states EOC exam scores are factored into students' course grades instead of or in addition to students being required to pass the exam to receive credit (Domaleski, 2011). Still, these exams are broadly based on the same theory of action that more traditional comprehensive exit exams are based on.

Early evaluations of HSEEs describe the motivations behind the policies (Chudowsky et al., 2002). One overriding motivation is to provide some assurance that the high school diploma “means something”—that it signifies that the diploma recipient has indeed demonstrated sufficient attainment of some body of knowledge and skills needed to succeed beyond high school. By establishing clear expectations and ensuring that students must meet those expectations to receive the diploma, the HSEE should ensure at least a minimum standard has been obtained. If the standard is sufficiently rigorous, another motivation for HSEEs is that they could boost the rigor or quality of the education students are provided (i.e., setting a baseline beneath which high school teaching and learning will not fall). Third, HSEEs are seen as motivating individual students to work harder, and indeed there is evidence that high school students in the United States have not felt sufficiently challenged (Boser & Rosenthal, 2012). Fourth, HSEEs that are well aligned to state standards can provide additional weight behind standards-based reform policy, driving high schools to better align instruction to standards. Fifth, these exams could have formative uses if their results are used to identify students falling below the target and provide them support or remediation to succeed. And last, HSEEs are seen as possibly strengthening college applicant pools or providing a new source of assessment data that is better aligned with the high school curriculum than are the SAT and ACT. Advocates also note that many competitor countries in Europe and Asia have high-stakes student-level exit exams.

Grade Promotion Exams

A second main type of assessment for student accountability is the grade promotion exam. Grade promotion exams are used by states and districts to hold individual students accountable for meeting performance thresholds to be promoted to the next grade. The most common type of grade promotion exam is the third-grade reading exam, which was used by 17 states and the District of Columbia as of the 2018–2019 school year (National Conference of State Legislatures, 2019). Third-grade reading is seen as especially consequential for students' future success, with evidence that the large majority of students who eventually drop out were struggling readers by the time they were in third grade (D. J. Hernandez, 2012). Eight other states allowed the use of state tests for third-grade retention purposes but did not require it. And there are a number of large urban districts that now use or have in the past used third-grade reading

assessments for this purpose (e.g., Chicago, New York). Importantly, there are very few states or localities that do not offer numerous exemptions to the test-based promotion rules. These so-called good-cause exemptions include students with limited English proficiency (typically 3 or fewer years in an English language acquisition program); special education students; participating in an intervention; parent, principal or teacher recommendations; previous retention; demonstrating proficiency through a portfolio (student work demonstrating mastery of academic standards in reading); or passing an approved alternative reading assessment. While well intentioned and aligned with best practices in test use, recent research found that these exemptions are inequitably applied, with children of more educated mothers substantially more likely to obtain exemptions (LiCalsi et al., 2019).

There are also some test-based promotion policies at other grade levels, though it is harder to get an accurate count on the prevalence of these kinds of policies. For instance, Chicago's test-based promotion policy that began in the 1990s included promotion tests in third, sixth, and eighth grades (Jacob & Lefgren, 2009; Roderick & Nagaoka, 2005), and New York City's included Grades 3 through 8 (Mariano & Martorell, 2013).

The theory of action behind grade promotion exams is typically twofold. First, grade promotion exams are generally tied to specific remediation and intervention programs, such as summer school, assignment to a high-quality teacher, or extended reading instruction (Ozek, 2015; Winters & Greene, 2012). Through this path, the exam identifies students who need additional support, and the policy responds by providing that support. If the support is effective, student abilities and skills will increase, producing long-lasting effects. Second, like exit exams, promotion exams are thought to have motivational effects on students. Indeed, a common complaint from teachers about most state tests under standards-based accountability is that students do not have much incentive to try hard because there are few consequences for individual students; promotion exams address this concern and may also cause students to work harder in their learning. We discuss the literature on the impact of these student accountability systems in the section "The Impact of Accountability Policies".

TEACHER ACCOUNTABILITY POLICY SYSTEMS

A relatively recent target of test-based accountability is individual teachers. The Obama administration's Race to the Top and ESEA flexibility waiver initiatives dramatically changed the way that teachers around the nation were evaluated. These changes were in response to two streams of research. First, a variety of studies provided compelling evidence that teachers have large effects on students' achievement (e.g., Rivkin et al., 2005; Rockoff, 2004). Subsequent research has deepened this literature and also extended it to teacher impacts on nontest measures (e.g., Blazar & Kraft, 2017; Chetty et al., 2014a; Jackson, 2018). Second, academic research (Almy, 2011; Donaldson, 2009; Sartain et al., 2011; Sinnema & Robinson, 2007; Stronge & Tucker, 2003; Toch & Rothman, 2008; Tucker, 1997) and one high-profile report (Weisberg et al., 2009) suggested that

teacher evaluation systems largely failed to differentiate teachers or to provide useful feedback to improve their teaching.

As a result of Race to the Top and the waivers, states substantially redesigned their teacher evaluation systems. The main features of these revised teacher evaluation plans included (a) the use of multiple measures of performance to evaluate teachers, (b) the inclusion of student outcomes (measured most often by state test scores or student learning objectives) in evaluation, (c) the use of three or more levels of ratings (i.e., not just “*Pass/Fail*”), and (d) the attachment of consequences to results of evaluations (Steinberg & Donaldson, 2016).

In terms of the component measures of new teacher evaluation systems, by far the most common measure—appearing in every state and district plan analyzed in a 2016 survey of 50 states and 25 large districts—was classroom observations (Steinberg & Donaldson, 2016). Observation-based ratings of effectiveness also comprised the largest portion of teachers’ overall evaluation scores, on average representing a bit more than half of the total rating. About one fifth of states and half of large districts also included a separate, but very lightly weighted (about 2%–5%, on average) measure of teachers’ professional conduct. Measures of student learning were also an essential part of teacher evaluation systems postwaiver/Race to the Top (RTTT). Test-based measures—value-added models or student growth percentiles, which we discuss later—were part of 80% of states’ and large districts’ evaluation systems. In other states and districts, or in nontested grades, the most common measure of student performance was student learning objectives. Finally, some districts and states included other measures in teacher evaluation systems. For instance, 26%–30% of states and districts included schoolwide achievement data in individual teachers’ evaluations, and about one in six states and districts included results from student surveys.

There were a number of important technical challenges in the design and implementation of these systems. One of the major challenges was in including measures of student learning that could be applied to all teachers. The large majority of teachers do not teach in subjects and grades for which student growth models can be calculated from state tests (i.e., all teachers below Grade 4 and above Grade 8, teachers of any subject other than mathematics or ELA). For these teachers, alternative measures of student learning could sometimes be computed. But these could sometimes result in seemingly nonsensical and indefensible measures being used, such as the art teacher who is evaluated in part by the growth in students’ mathematics achievement (Jacobs, 2015). Another common approach for these teachers was individualized “student learning objectives,” which were widely used but lacked standardization and adherence to psychometric standards (Buckley, 2015). They also ultimately do not seem to have resulted in meaningful instructional improvement (e.g., Lachlan-Haché, 2015).

Research has begun to shed light on the implementation and effects of these teacher evaluation policies. First, just as before these systems were implemented, very few teachers seem to be rated as needing improvement—a study of 24 state evaluation systems found fewer than 5% of teachers rated as needing improvement (Kraft & Gilmour,

2017). Second, the low percentage of teachers rated as needing improvement does not comport with district leaders' subjective ratings of teacher performance (i.e., when asked directly, leaders rate far more than 5% of teachers as *Below Proficient*; Kraft & Gilmour, 2017). Third, teachers' overall ratings in evaluation systems are highly sensitive to these systems' design decisions (Steinberg & Kraft, 2017). For instance, the greater the weight on student achievement and other norm-based measures of performance, the larger the proportion of teachers deemed *Not Proficient*. Finally, where there is a large literature on achievement-based measures of teacher performance, researchers have argued that relatively less is known about the properties of nontest measures of performance, such as teacher observations (J. Cohen & Goldhaber, 2016).

What we do know about nontest measures suggests that they may have many of the same challenges as test-based measures. For instance, there is evidence that the attachment of stakes to evaluation scores can change the distribution of those scores (Liu et al., 2019) and that there are racial gaps in teacher effectiveness ratings that are driven by school-level demographic differences (Steinberg & Sartain, 2021). Teacher observation scores can have substantial measurement error unless they are averaged over multiple lessons and raters (Ho & Kane, 2013). However, findings from randomized studies also suggest that observation scores can be unbiased estimates of instructional quality (Bacher-Hicks et al., 2017). Summarizing the impact of the Obama-era push for teacher evaluation reform, a recent national study found precisely zero effect on student achievement nationwide (Bleiberg et al., 2024).

THE IMPACT OF ACCOUNTABILITY POLICIES

What have been the effects of accountability on key teacher and student outcomes? This section reviews the available literature on these impacts, focusing on the effects on (a) student achievement as measured by standardized tests; (b) longer term outcomes, including graduation, college, and outcomes beyond college; (c) teacher outcomes, including their instruction; and (d) unintended consequences. This section draws on several recent reviews of accountability and its effects (Figlio & Loeb, 2011; National Research Council, 2011; Polikoff & Korn, 2020). First, we discuss the evidence focused on school accountability systems and then teacher and student accountability.

Effects of School Accountability on Student Achievement

There are several main methods that have been used to estimate the impact of school accountability policies on student achievement on state tests. The most prominent studies have used state-level data to compare changes in achievement levels and trends between states that implement new accountability systems and those that already had existing systems in place (using difference-in-differences or comparative interrupted time series methodologies). Investigating the impact of NCLB on student achievement, two studies (i.e., Dee & Jacob, 2011; Wong et al., 2015) used analyses of this type on data from NAEP. Both studies found evidence of positive to null effects, depending

on grade and subject. Dee and Jacob found achievement impacts of around a quarter of a standard deviation in elementary mathematics. Wong and colleagues found similar impacts in elementary mathematics, as well as statistically significant impacts in middle school mathematics and marginally significant impacts in elementary reading. Interestingly, these positive effects come despite the fact that NCLB's consequences were largely ineffective—supplemental education services saw low participation and ineffective program design (Heinrich et al., 2010), and vanishingly small numbers of students participated in public school choice, often because there were no options available (Lauen, 2008). Earlier studies similarly leverage variation among states in either the timing or the strength of accountability in attempts to isolate a causal effect; these studies generally find positive achievement effects of similar magnitude (e.g., Carnoy & Loeb, 2002; Hanushek & Raymond, 2005).

Another common approach to estimate the impact of accountability is to leverage within-state variation in state or district policies. Figlio and Loeb (2011) grouped these studies into three categories: (a) studies that look within states or districts over time (e.g., Jacob, 2005; Klein et al., 2000); (b) studies that compare districts that implement accountability policies to those that do not (e.g., Ladd, 1999; S. S. Smith & Mickelson, 2000); and (c) studies that examine the differential responses of schools subject to and not subject to accountability pressure using regression discontinuity design (e.g., Rockoff & Turner, 2010). They concluded that there is evidence of a positive relationship between school accountability and student achievement, but that this relationship is not universal and varies in magnitude.

The designs of most state accountability systems are well suited to the use of regression discontinuity designs, with schools just above and below the threshold functionally similar except for their accountability classification. Studies in North Carolina (Ahn & Vigdor, 2014), Wisconsin (Chakrabarti, 2014), and Kentucky (Bonilla & Dee, 2020) found positive effects of accountability on student achievement. In contrast, Dee and Dizon-Ross (2017) found null effects in Louisiana, Hemelt and Jacob (2020) found essentially null effects in Michigan, and Atchison (2020) found null to negative effects in New York State. A study of New York City school grades by Winters and Cowen (2012) also used regression discontinuity (RD) designs at the C/D and D/F thresholds. That study found heterogeneous effects, with positive impacts at the D/F margin but negative impacts at the C/D margin.

Studies have also investigated the heterogeneous effects of accountability policies on different types of students, often with a focus on the impact of accountability on student achievement gaps. This is an important area of research, because NCLB and other accountability policies generally aim to close racial and socioeconomic achievement gaps. One study of achievement gaps in North Carolina found that NCLB accountability narrowed the Black–White test score gap by around 4%–10% of the baseline gap in both mathematics and reading (Gaddis & Lauen, 2014). Other studies also provided evidence of achievement gap effects by showing that accountability affected subgroup achievement differently, again generally leading to a modest narrowing of gaps (e.g., Dee

& Jacob, 2011; Figlio et al., 2009; Lauen & Gaddis, 2012). Pre-NCLB accountability studies found mixed effects, with Hanushek and Raymond (2005) finding a narrowing of the White–Hispanic gap and Harris and Herrington (2004) finding no impact of accountability on performance gaps.

None of the above-cited studies is generalizable unconditionally, but the results taken together demonstrate that school accountability policies often have a positive impact on average student achievement. These effects may be concentrated in mathematics and in the early grades. And the effects on achievement gaps are modest, at best. The impacts of accountability appear not to be uniform, however, because there is evidence that these effects are stronger for the lowest-performing schools (Figlio & Rouse, 2006; Jacob, 2005), in states with higher proficiency standards (Wong et al., 2015), and in states with greater degrees of local autonomy (Loeb & Strunk, 2007).

Our conclusion about the impact of accountability aligns with that of Figlio and Loeb (2011), who said, “Taken as a whole, the body of research on implemented programs suggests that school accountability improves average student performance in affected schools, at least in general” (p. 410). And our conclusion is more positive than that of the National Research Council (2011), which noted,

Test-based incentive programs . . . have not increased student achievement enough to bring the United States close to the levels of the highest achieving countries. When evaluated using relevant low-stakes tests . . . the overall effects on achievement tend to be small and are effectively zero for a number of programs. Even when evaluated using the tests attached to the incentives, a number of programs show only small effects. (p. 4)

The differences between our conclusion and that of the National Research Council may have to do with at least two factors. First, the literature on this topic has grown dramatically in recent years, with many of the positive effects coming in more recent studies using more advanced methods. And second, the difference hinges on the definition of “small” as pertains to the magnitude of accountability effects. We agree with Kraft (2020) that effects of even just .2 standard deviations in performance are not small, even though they pale in comparison to the magnitude of performance gaps and are clearly insufficient to narrow gaps with higher performing nations. This is especially the case insofar as the effects were cumulative across grades and years—the resulting long-term impacts could indeed be quite large.

Despite the positive evidence overall, there appears to be very modest evidence of any significant gap-closing effects of accountability systems. This is notable insofar as student group accountability is intended to drive educators to work to narrow these gaps—that portion of the theory of change clearly appears to be failing. Furthermore, concerns about accountability policies are often driven by equity issues—that standards-based accountability undermines possibilities for culturally relevant pedagogy (Royal & Gibson, 2017), leads to proceduralized instruction in classrooms serving historically marginalized youth (Polikoff & Struthers, 2013), and overall undermines

educational opportunity (Rebell, 2008). At the very least, it is by now clear that performance gaps will need to be closed with other policy levers rather than through test-based accountability.

Effects of School Accountability on Longer Term Student Outcomes

Much less is known about the long-term effects of accountability or the effects on outcomes besides student academic achievement. In large part, this is due to data limitations that may be addressed as new state longitudinal data systems mature. There are two exceptions to this pattern. A study using longitudinal graduation rate data from before and after the NCLB era found that the introduction of accountability systems coincided with a significant bump in high school graduation rates equal to about one fifth of the descriptive increase over that time (Princiotta, 2019). A second study, by Deming et al. (2016), explored the effect of Texas high school accountability policies during the 1990s on college attainment and future earnings. They found that schools on the threshold of a *Low Performing* label increased student achievement on high-stakes exams, which led to higher rates of college enrollment and completion and higher salaries at age 25. In contrast, they found a negative effect of accountability in higher performing schools, which they attributed to the reclassification of low-performing students into categories exempt from taking state tests (e.g., special education). There is a great need for further research on the impact of accountability on longer term outcomes.

Effects of School Accountability on Instruction and Teacher Outcomes

The theory by which accountability policies boost student achievement operates through these policies' effects on teacher behaviors. The effect of accountability on teacher practice and other teacher outcomes is even more difficult to estimate. This is primarily because there is a lack of high-quality measures of teachers' instruction at the kind of scale needed to conduct causal analyses. However, there are many correlational and descriptive studies that explore teachers' instructional responses to school accountability, especially focusing on teachers' instructional alignment to standards. The results of this work generally suggest modest instructional change in the direction of the standards, as called for by the theory underlying standards-based accountability.

A large number of self-report studies using surveys or interviews indicated that teachers believe they practice aligned instruction and are making efforts to increase the alignment of their instruction with standards and assessments (see, for instance, Hamilton, 2012; Pedulla et al., 2003). However, when teachers' practice is assessed by external means, such as through observation, it often appears less aligned to the intent of standards than teachers think (see, for instance, D. K. Cohen, 1990; H. C. Hill, 2001; McDonnell, 2004; Spillane, 2004). Still, there is evidence that the quality of instruction

and school goals for student performance have improved over the early years of the accountability era (Hunter, 2019; Lee & Lee, 2020).

A different approach to studying alignment involves surveying teachers about the content of their instruction and then using alignment methods (discussed in more detail in the section “Methods of Measuring Alignment”) to compare these survey responses to content analyses of standards and assessments (see Polikoff, 2012a; Porter, 2002). These approaches make it less likely that social desirability and other self-report issues could bias the results. Several studies that use these techniques find that: (a) the alignment of instruction to standards is low, much lower than reported on other kinds of surveys; (b) instructional alignment generally increases over time, especially in mathematics; and (c) some state policy features and characteristics of teachers predict stronger instructional alignment (Polikoff, 2012a, 2012b, 2013). Still, the findings taken together indicate significant gaps in the extent to which teachers are implementing standards in the classroom, even decades into the standards era and with considerable accountability pressure.

There is also some evidence that accountability affects various dimensions of teachers’ working conditions, though these effects seem modest and mixed in direction. Two studies of NCLB, for instance, found contrasting effects. Grissom et al. (2014) found that the law had no impact on teachers’ hours worked, job satisfaction, or commitment to teaching; had small negative effects on teachers’ perceptions of cooperation; and had small positive effects on teachers’ perceptions of classroom control and administrator support. Dee et al. (2013) found that the law increased teachers’ salaries and teacher-reported student engagement. Two other studies have found that accountability decreased teachers’ job security (Reback et al., 2014), increased the frequency of involuntary school transfers, and decreased the frequency of involuntary attrition (Sun et al., 2017). And a more recent study on just the impact of grade-level accountability testing requirements found no impact on teacher attrition (Fuchsman et al., 2020). In sum, the effect of accountability on teacher working conditions appears mixed and modest.

Unintended Consequences of School Accountability

As indicated in Figure 16.1, there are also a number of well-documented unintended consequences of accountability policies, which arise when teachers and schools strategically seek to maximize their scores and avoid accountability consequences. One type of gaming behavior is through the reallocation of instructional time. For example, schools have been shown to allocate instructional time away from nontested subjects and toward tested ones (Dee et al., 2013; Judson, 2013). Another variant on this is the allocation of instructional time to specific test-taking strategies (for instance, using class time to take practice tests, teach strategies for completing specific types of test items, or narrowly focus on content that is expected to be assessed; Jennings & Bearak, 2014; Jennings & Sohn, 2014). There is ample evidence that these behaviors have taken place under recent accountability regimes. Thus, positive effects that are measured on state

tests should generally be cross-validated with lower stakes exams like NAEP (Koretz, 2008).

Another type of strategic manipulation is when educators narrowly target resources and instructional efforts on the students who matter the most for accountability ratings. Under NCLB, for example, school performance was determined by the percentage of students scoring above a proficiency cut point on the state test. This creates incentives to target efforts on students just below the proficiency cut point (often referred to as bubble kids), since improving those students' scores would be more likely to boost a school's performance rating. There is some evidence that teachers diverted attention to bubble kids (Booher-Jennings, 2005; Jennings & Sohn, 2014), including evidence that positive effects of accountability were concentrated on students close to the proficiency cut score (Neal & Schanzenbach, 2010). However, other studies have not replicated the bubble kids finding using achievement data (e.g., Cronin et al., 2005).

The most damaging and pernicious unintended consequence of accountability policies is perhaps their effect on teacher cheating, which can be enabled or even encouraged by school and district leaders. High-profile cheating scandals have emerged in some places, and these scandals have often been blamed on the intense pressure associated with accountability policies. For example, an Atlanta cheating scandal ended in 11 teachers being convicted on racketeering charges (Blinder, 2015). A study in Chicago used unusual test score patterns and identified that approximately 5% of teachers had likely cheated on the state exams (Jacob & Levitt, 2003). Cheating-type behaviors have also been reported in nontest outcomes, such as graduation rates (Edwards & Mindrila, 2019).

Effects of Student Accountability

Exit Exams

The literature on the impact of HSEEs on student achievement, high school graduation, and other desired outcomes consistently indicates that these exams do not lead to the outcomes desired by their proponents (Holme et al., 2010). At the same time, the worst fears of HSEE opponents—that they would lead to large-scale reductions in opportunity, especially for disadvantaged groups—appear not to have materialized given that graduation rates have trended upward for decades and, prepandemic, stood at all-time highs overall and for all racial-ethnic groups (Harris et al., 2020).

A number of studies using quasi-experimental designs have examined the impact of HSEEs on student achievement (Grodsky et al., 2009; Jacob, 2001; Reardon et al., 2010), graduation/dropout rates (O. Baker & Lang, 2013; Dee & Jacob, 2006; Hemelt & Marcotte, 2013; Ou, 2009; Papay et al., 2010; Polson, 2018; Reardon et al., 2010), and labor market outcomes (Baker & Lang, 2013; Warren et al., 2008). In general, these studies have been focused on comprehensive exit exams rather than EOC exams. Some studies estimate the causal impact of just passing the exam compared to the counterfactual of just failing. Other studies estimate the causal impact of a policy year that includes an exit exam compared to the counterfactual of a policy year that does not.

Broadly speaking, this literature produces several clear conclusions (for a comprehensive review of evidence on exit exams, see Holme et al., 2010). First, exit exams seem to have no effect on student achievement as measured by standardized tests (Grotsky et al., 2009; Jacob, 2001; Reardon et al., 2010). Second, exit exams do modestly decrease graduation rates and increase dropout rates, and these effects are concentrated among Black students and very-low-performing students and in states that have more rigorous exams (Baker & Lang, 2013; Hemelt & Marcotte, 2013; Jacob, 2001). And third, these exams have no appreciable effect on labor market outcomes (Baker & Lang, 2013; Warren et al., 2008). In response to these disappointing findings, there was mounting pushback against these exams starting in the mid-2000s. This pushback followed many researchers who opposed these exams from their onset; for a discussion see, for example, Warren & Grotsky (2009).

Promotion Exams

There is a large literature on the impact of test-based promotion policies on a range of short- and long-term outcomes for students, including student achievement in the tested grade (Winters, 2018); subsequent achievement and promotion (Greene & Winters, 2007; Mariano & Martorell, 2013; Mariano et al., 2018; Schwerdt et al., 2017); attendance, graduation, and other attainment measures (Eren et al., 2017; Jacob & Lefgren, 2009; Mariano et al., 2018; Schwerdt et al., 2017); and other outcomes like behavior, incarceration, and college enrollment (Eren et al., 2018; Ozek, 2015; Schwerdt et al., 2017). In general, this literature uses regression discontinuity methods to provide convincing causal estimates of the impact of these policies. However, there are methodological challenges in this work, especially in terms of choosing an appropriate comparison group, tracking impacts longitudinally, and separating out the effects of the retention policy from other portions of the intervention, like summer school prior to the retention decision (see Roderick & Nagaoka, 2005, for a discussion of these and other challenges).

In terms of short-term effects, studies found mixed evidence that the threat of retention induced by the policy boosts student achievement, with some evidence that it does, especially for the lowest-performing students (Winters, 2018). Subsequent to the retention decision, research suggests there are positive and substantial effects of retention on achievement, but that these effects depend on who retained students are compared against (same-grade vs. same-age peers; see Mariano & Martorell, 2013, for a good discussion of this issue; Greene & Winters, 2007; Schwerdt et al., 2017; Winters & Greene, 2012). There are some studies showing smaller positive or null effects (Jacob & Lefgren, 2004; Roderick & Nagaoka, 2005) and there is also evidence that test score effects may diminish over time (Schwerdt et al., 2017; Winters & Greene, 2012). Test-based retention policies seem to have null or perhaps modest negative effects on long-term outcomes like high school completion (Eren et al., 2017; Jacob & Lefgren, 2009; Mariano et al., 2018; Schwerdt et al., 2017), with evidence that retention policies that occur later in students' schooling may be especially harmful for the completion of the lowest-performing students (Jacob & Lefgren, 2009; Mariano et al., 2018). Research

on other outcomes found that test-based promotion policies can lead to increases in adult crime and student misbehavior (Eren et al., 2017, 2018; Ozek, 2015).

In short, the literature suggests that whatever benefits there may be from test-based retention and support in the early years, the long-term impacts of these policies appear to be close to zero and may in some cases be negative. These modest findings are notable given the large cost of student retention and its disproportionate effects on historically marginalized student groups (West, 2012). Importantly, the effects of test-based retention policies undoubtedly vary as a function of the rigor of the assessment and cut score, the nature of the interventions and supports that are paired with test failure, and the extent to which there are nontest alternatives available for students.

Special Considerations for High-Stakes Student-Level Accountability

In their review of high-stakes testing for student placement and promotion, Heubert and Hauser's National Research Council report (1999) identified three criteria for fair test use for students: psychometric adequacy, opportunity to learn, and educational benefit. The *Standards* (AERA et al., 2014) also have specific criteria listed for these tests in an illustrative example:

When tests are used for promotion and graduation, the fairness of individual score interpretations can be enhanced by a) providing students with multiple opportunities to demonstrate their capabilities through repeated testing with alternative forms or other construct-equivalent means; b) providing students with adequate notice of the skills and content to be tested, along with appropriate test preparation materials; c) providing students with curriculum and instruction that afford them the opportunity to learn the content and skills to be tested; d) providing students with equal access to disclosed test content and responses as well as any specific guidance for test taking (e.g., test-taking strategies); e) providing students with appropriate testing accommodations to address particular access needs; and f) in appropriate cases, taking into account multiple criteria rather than just a single test score. (p. 186)

Unlike Heuber and Hauser (1999), the *Standards* (AERA et al., 2014) do not count educational benefit to the marginal or average student as a criterion. As evidence continues to mount that these student-level accountability exams have no long-term positive effect for average or marginal students and are sometimes detrimental to historically marginalized groups (Penfield, 2010), the burden of proof should be considerable for current and new programs to argue that existing or proposed remediation strategies for failing students have a positive causal effect over social promotion. As we reviewed in an earlier section, tests serve political purposes as well as educational purposes, including signaling worthy goals and managing and controlling systems. It would be best if these purposes could be served with minimal detriment to students, including by following the guidance from the *Standards* and opening up multiple pathways to demonstrating readiness.

DERIVED ACCOUNTABILITY SCORES WITH A FOCUS ON GROWTH MODELS

As Table 16.1 indicates, the type of score creates opportunities to both achieve intended aims and inspire unintended negative consequences in accountability systems. The primary NCLB reporting metric was the percentage of proficient students, an aggregate status measure that was an obviously poor proxy for school quality because of its conflation with prior learning and other out-of-school opportunities to learn (Linn et al., 2002; Ryan, 2004). Proficiency percentages also exhibit deeply problematic properties when they become the basis of score trends and gap trends, where any group with proficiency percentages near 50% is expected to manifest greater trend magnitudes because of the density of the distribution local to the proficiency cut score (Ho, 2008; Holland, 2002). As we reviewed in previous sections, the statistic also enables gaming behavior related to triaging low-scoring students who may require greater effort to teach to proficiency. And it results in a considerable loss of information about differences and changes in student performance at other locations of the score scale. In these ways, the selection and properties of the derived score for any accountability purpose deserve scrutiny. In this section, we illustrate this by discussing the use of growth metrics for accountability purposes.

A debate about “status versus growth” motivated flexibility in the form of the GMPP (2005). In the GMPP announcement, Secretary of Education Margaret Spellings never defined what “growth” meant (U.S. Department of Education, 2005). The announcement insisted only that models adhere to seven “bright line principles,” such as “Ensure that all students are proficient by 2014,” and that the model “must track student progress.” Subsequent guidelines from the peer review panel similarly left latitude for growth definition and model specification (U.S. Department of Education, 2006).

The guidelines at the announcement of the Race to the Top competition seemed to be more explicit and defined growth as “the change in achievement data for an individual

Table 16.1 An Example of a Categorical Model From Delaware’s 2009–2010 School Year

	Year 2 Level				
Year 1 Level	Level 1A	Level 1B	Level 2A	Level 2B	Proficient
Level 1A	0	150	225	250	300
Level 1B	0	0	175	225	300
Level 2A	0	0	0	200	300
Level 2B	0	0	0	0	300
Proficient	0	0	0	0	300

student between two or more points in time” (U.S. Department of Education, 2009, p. 59742). However, it then continued, “A State may also include other measures that are rigorous and comparable across classrooms,” along with its motivations “to allow States the flexibility to develop data and assessment systems” (p. 59742). This left space for GMPP models to continue and effectively took no position on the definition of growth, a degree of flexibility that continues under ESSA.

As the final report of the GMPP described (Hoffer et al., 2011), states took a variety of approaches to operationalizing growth. Models continue to proliferate through the ESSA era, including renewed focus on “through-course” or “through-year” models that track growth from a fall test through a spring test. In this section, we review prototypical models and demonstrate that these models operationalize growth using related but fundamentally distinguishable approaches, and differently at the student level than at the aggregate level.

The proliferation of growth models in a policy space constructed deliberately to allow for flexibility has led to confusion among terms and definitions. We follow the general framework and nomenclature provided by Castellano and Ho’s *A Practitioner’s Guide to Growth Models* (2013a). In their guide, Castellano and Ho attempted to be explicit about each growth model and included its aliases and statistical foundations. In addition, they articulated the primary interpretations that models support in terms of three answerable questions: how much growth, growth to where, and what caused growth?

The gain-based model is an intuitive and largely straightforward model that requires a vertical scale, a challenging constraint with stringent requirements reviewed in Moses (this volume). The categorical model is a flexible framework that considers student status in a small number of categories (usually four to nine) and operationalizes growth in terms of transitions between categories (R. Hill et al., 2006). And conditional status models, including the student growth percentile model (SGP; Betebenner, 2009), express growth in terms of status beyond expectations given past scores. Value-added models are related to conditional status models and associate aggregate conditional status with the value that teachers and schools add to test scores.

Gain-Based Models

The first and arguably most intuitive growth model is a trajectory that each student has over time that can be described by a slope or a gain. The statistical foundation is the simple difference between two scores. Extensions are straightforward and can include estimating trajectories over more than two points in time or allowing for non-linear trajectories, in the tradition of longitudinal data analysis (e.g., Singer & Willett, 2003). Although gain scores are intuitive, they become problematic in that they rely on vertical scaling decisions, whereby expected gains may differ in magnitude across grades. Although average gains may be an accurate representation of the amount of learning in each grade on an absolute scale, this may be more attributable to the typical developmental trajectories of children than to schools. Straight comparisons of gains

on developmental vertical scales are thus inappropriate for comparing the amount of growth for which teachers and schools may be responsible.

Categorical Models

The categorical model, also known as the value table or the transition matrix model, divides each within-grade score scale into a smaller number of ordered categories (R. Hill et al., 2006). Table 16.1 shows Delaware's value table for the 2009–2010 academic year (Delaware Department of Education, 2010). A student who scores in Level 1A in Year 1 but Level 2A in Year 2 receives a growth score of 225, as shown, and the average across students in the school represents the school-level score. The model relies more than others on the selection of cut scores, where transitions between categories function as student growth data. Logically, the cut scores between Level 1B and Level 2A must have some basis for equivalence.

The categorical model is flexible in the sense that values for particular transitions between categories can be adjusted to user specifications. As R. Hill et al. (2006) demonstrated, careful selection of values for particular transitions can result in a pre-growth-era status model, where only proficiency is counted, or something that seems more gain based, where the gain is quantified as the number of levels that are gained or lost. The cost of this flexibility is the loss of information that comes with categorization, where the model cannot distinguish between the very highest and the very lowest scores in any given category. Although this may seem inappropriate for comparing growth for individual students, at the aggregate level, the errors due to coarse categorization are diminished, particularly as the number of categories increases. However, at a certain number of categories, judgments that support differing values become more difficult to distinguish and justify, and the model becomes likely to reduce to something similar to a gain-based model, with a number of categories approaching the number of score points.

The categorical model technically provides growth descriptions. However, the values that are selected for the categorical model may be motivated by inferences about whether a particular transition between categories is sufficient to warrant an “on-track” designation for students making that transition. To the extent that these inferences inform the choice and interpretation of values, the function of a categorical model is one of growth prediction, as well as growth description. In the case of some growth models, like Iowa's model under its GMPP (Hoffer et al., 2011), this took the form of values like those in Table 16.1, except any nonzero value was simply a 1. This was based in part on the argument that a gain in categories established students as being on track to *Proficient*. In this way, the categorical model can support both growth descriptions and growth prediction.

Conditional Status Models and the SGP Model

Betebenner (2009) introduced the SGP metric as a normative approach to describing student growth. The SGP metric uses nonlinear quantile regression to support

conditional status interpretations, where the current status of a student is referenced to expected percentiles given the score history of students. Although the name of the metric seems to indicate the percentile rank of a student's gain score, the statistical foundation is regression based, where a student's current status is considered in light of expectations given past scores.

Castellano and Ho (2013b) reviewed the SGP estimation procedure in detail. Fitting the statistical model can be time-consuming for large data sets and uses an open-source R library (Betebenner et al., 2023). The SGP is calculated by first estimating 100 nonlinear quantile regression manifolds, for quantiles from .005 to .995, where the outcome variable is the "current" score and the predictor variables are all prior-year scores. Castellano and Ho (2013b) demonstrated that this is practically similar to a straightforward linear regression model of current-year scores on past-year scores, where the SGP corollary is the percentile rank of residuals. The school-level SGP metric used most often in practice is the median SGP, which Betebenner (2008) argued for on the basis of the ordinal nature of percentile ranks. Castellano and Ho (2015) showed that means are superior, where medians require around three times the sample size for the same precision.

Value-Added Models for Teachers and Schools

As the focus in the 2000s shifted from proficiency to growth, value-added models (VAMs) rose as a candidate approach that unabashedly attempted to estimate the causal effect of a teacher or school on test scores, compared to a counterfactual average teacher or school. One set of questions concerned whether and how to estimate these causal effects, and another set of questions concerned whether and how to use them to hold teachers and schools accountable to student learning. High-profile consensus reports argued against (E. L. Baker et al., 2010) and for (Glazerman et al., 2010) the use of VAMs for evaluating teachers. The American Statistical Association (2014) offered a more neutral set of cautions that was rebutted by economists (Chetty et al., 2014b). And the American Educational Research Association (2015) listed limitations and specified eight technical requirements. We briefly review some of the technical issues here and reserve policy models for teacher and school evaluation for the next section.

Koedel et al. (2015) reviewed the substantial evidence about VAMs that emerged in this period. Using experimental evidence where students are actually assigned randomly to teachers, Kane and Staiger (2008) and Kane et al. (2013) estimated experimental causal effects and showed that nonexperimental VAM estimates are statistically indistinguishable from experimental estimates. Bacher-Hicks et al. (2014) and Chetty et al. (2014a) used a quasi-experimental approach to estimate causal effects of teachers from successive cohorts that have undergone staffing changes. They also found that VAM approaches recover these quasi-experimental estimates well. The upshot is that nonexperimental methods are good approximations of quasi-experimental and experimental results in the few studies we have. Further evidence of bias when test scores are high stakes would be useful.

Koedel et al. (2015) also reviewed the substantial evidence about VAM imprecision and intercorrelation, where the general finding is that VAM is positively but weakly correlated across subjects and years (Corcoran et al., 2011; Goldhaber et al., 2013; Lefgren & Sims, 2012; Lockwood et al., 2007). VAM estimates are also positively but weakly correlated across high-stakes versus low-stakes tests (Corcoran et al., 2011; Papay, 2011) and positively but weakly correlated with other evaluation measures like principal, peer, and student ratings (Kane & Staiger, 2012).

Koedel et al. (2015) established sensible technical recommendations for VAMs, including the use of student and school covariates, free estimation of coefficients for prior-year scores (instead of fixing coefficients to 1 to fit gain-based models), and corrections for measurement error (Lockwood & McCaffrey, 2014). They also noted that fixed-effects models (one-step estimation) are similar to averaging residuals (two-step estimation), with the latter approach benefiting from computational feasibility and separate inclusion of classroom and school covariates.

Suitability of Growth Metrics for Accountability Purposes

In test-based accountability systems, there are trade-offs among different growth models along dimensions of transparency, accuracy, and the potential for negative consequences (Ho, 2014). Gain-based models will seem transparent to users for individual reporting but create inequities in accountability models as a result of dependencies between initial status and expected growth. Categorical models may seem transparent to designers of accountability systems but reduce precision as a result of coarsening and confuse point-based aggregate policy goals with score-based individual learning goals. For student-level reporting, SGPs may seem like a black box, whereas at the aggregate level there may be insufficient control of sociodemographic variables for high-stakes purposes. And VAMs may be similarly imprecise and create incentives to artificially deflate test scores from prior years. In these ways, growth models illustrate the necessity of attending to the full complexity of Figure 16.1 by specifying the level of action, score type, and intended purposes, while anticipating predictable unintended consequences.

EVALUATING THE ALIGNMENT OF ASSESSMENTS FOR ACCOUNTABILITY

We next turn to a series of important technical issues underlying test-based accountability policies. There is no measurement concept more central to standards-based education policy efforts than alignment. Even the earliest versions of standards-based education reform in the 1990s had alignment at their core. (For instance, the text of the ESSA uses a form of the word “align” 72 times, demanding alignment among an array of policy instruments, including state content standards, state summative assessments, alternate assessments, English language proficiency assessments, and college entrance requirements.) A central goal of the standards movement is to get teachers to align their instruction with state content standards. To accomplish this, state policies

are meant to align challenging assessments with those standards and to support them through aligned curriculum materials and professional learning opportunities.

But what is alignment, and how has it been conceptualized and measured? This section draws on recent alignment reviews (Martone & Sireci, 2009; Polikoff, 2022) and discusses the role of alignment in accountability policy and the state of existing accountability assessments in terms of their alignment to standards.

Alignment and Validity

When assessments are to be used for accountability purposes, as has been required under federal law for several decades, it is imperative that those assessments align with the content standards they are intended to assess (and at the performance levels required) if the tests are to be used to gauge schools' contributions to student learning. This is certainly true from a policy design standpoint—aligned assessments send educators the message that the standards are important and should be well implemented, and they are at the heart of standards-based reform theory (M. S. Smith & O'Day, 1991). But it is also true from a validity standpoint—all of the inferences about student, teacher, and school performance that are made on the basis of assessment results rely to some degree on the extent to which those assessments align with the content standards they claim alignment to.

Consider the most common type of accountability use of assessments since the NCLB era—the rating of school performance based on its students' scores on state summative assessments. Setting aside policy design decisions about the construction of accountability metrics (e.g., status vs. growth and the use of nontest measures like attendance and graduation rates), the intended goal of these assessments is to gauge how effectively a school is educating children as to the content in state standards. Thus, to make valid inferences about school effectiveness, the assessments must measure student knowledge and understanding of the content and skills in those standards. Alignment is how we gauge the extent to which the assessments measure student knowledge and understanding of the standards. The argument is the same for teacher-level and student-level accountability assessments: To make valid inferences about teacher effectiveness or student mastery, the assessments must measure student knowledge and understanding of the standards.

Alignment is closely related to other terms—opportunity to learn, instructional validity, and instructional sensitivity. Opportunity to learn is a more general term that refers to “inputs and processes within a school context necessary for producing student achievement of intended outcomes” (Elliott & Bartlett, 2016). Alignment of instruction with standards can therefore be thought of as one component of opportunity to learn. For students to have the opportunity to learn the standards, they must (among other things) receive instruction that is aligned with the standards. For students to demonstrate their opportunity to learn the standards, the assessments must (among other things) be aligned to the standards.

Instructional validity and *instructional sensitivity* are other terms that have been used almost interchangeably with opportunity to learn. An assessment is instructionally

valid or instructionally sensitive to the extent that results on the assessment accurately reflect the content and/or quality of instruction students received (whereas some definitions of instructional validity only emphasize content, most implicitly cover both content and quality of instruction; see, e.g., D'Agostino et al., 2007). Presumably, a test could be aligned to standards and not instructionally valid (if, for instance, there were substantial gaps in the standards-aligned instruction students received). Instructional validity or sensitivity is more relevant to some uses of accountability tests than others (for instance, for a student promotion exam it might not matter how a student learned particular content, though it would matter tremendously that the student had the opportunity to learn it). For more on the distinctions among these terms, see Elliott and Bartlett (2016) and Polikoff (2010).

Methods of Measuring Alignment²

There are three main approaches to measuring alignment, but two of them are by far more widely used (Martone & Sireci, 2009). The most widely used procedure to investigate the alignment of state tests with state standards is the Webb approach. This approach compares tests with standards using four main criteria: categorical concurrence (agreement in coverage of broad content areas), depth of knowledge (agreement in coverage of depth or cognitive complexity), range of knowledge (breadth of coverage of the test relative to the standards), and balance of representation (even representation of content on the test). The Webb model also considers the source of challenge for test items (i.e., whether the source of challenge is construct relevant or whether the item is merely “tricky”). Expert raters analyze the content and cognitive complexity of each item on a test and use that information to calculate indices for each criterion. These indices are then compared with specified benchmarks to judge alignment along each criterion (see Webb, 1999, for more detail).

The second widely used alignment approach is based on the Surveys of Enacted Curriculum (SEC; Porter, 2002). Instead of directly comparing standards with assessments, this approach maps standards and assessments onto neutral content languages. These content languages, which define content at the intersection of specific topics (e.g., multiplying fractions) and cognitive demands (e.g., perform procedures), are intended to be exhaustive as to the content that would normally be taught in a particular grade and subject. Once multiple raters have coded the topics and cognitive demands in the standards and assessments, these codes are compared to one another by way of an alignment index that indicates the proportional agreement, as well as heat maps that graphically display areas of alignment and misalignment (Porter, 2002). Multiple alignment indices have been used in the literature (Polikoff et al., 2011). The SEC approach has also been extended to teacher surveys (e.g., Polikoff, 2012a, 2012b; Porter et al., 2011) and to curriculum materials (Polikoff, 2015), and similar alignment calculations have been done in those contexts.

The third most widely used approach was pioneered by Achieve to rate alignment of state tests with standards (Rothman et al., 2002). The Achieve approach rates the

individual test items on several criteria: confirmation of the test blueprint, the match of the content to its intended objectives, the match of the cognitive demand to what is called for by the objectives, the source of the item's challenge (whether it is the target objective or something construct irrelevant), and the overall level of cognitive demand. Then, it takes these item ratings and calculates holistic ratings for (a) level of challenge, (b) balance, (c) range, and (d) rigor (all but range are qualitative ratings; see Rothman et al., 2002, for a description of each dimension). Achieve differs from the previous two approaches by using a qualitative approach to holistic ratings—that is, relying on narrative summaries of the results rather than turning them into quantitative indices (Martone & Sireci, 2009).

There are also other approaches, which have been used more sporadically. For example, one approach has been recently created specifically for Common Core standards. This approach was used in a recent study of four assessment systems: the Partnership for Assessment of Readiness for College and Careers (PARCC), Smarter Balanced, ACT Aspire, and the Massachusetts Comprehensive Assessment System (MCAS; Doorey & Polikoff, 2016). This approach asks expert raters to analyze assessments against a large set of criteria that address both the content and the depth of the standards (for example, whether ELA items require students to read closely and provide evidence from the text, whether they use a balance of high-quality literary and informational texts, and whether they require a range of cognitive demand). These ratings are then rolled up to the whole test to indicate whether the assessment has a weak match, limited/uneven match, good match, or excellent match to the standards. This methodology has only been used in one study as of this writing, but it may be appropriate for more studies if the Common Core State Standards (CCSS) or similar standards remain in place. Other alignment approaches are summarized by Forte (2017).

There are strengths and weaknesses to each of these alignment approaches, and they are described in detail by Forte (2017). Some of the relevant distinctions include the following:

1. All of the methods are dependent on expert raters and are therefore time and labor intensive. The alignment ratings and their consistency across studies have sometimes been problematic (Herman et al., 2007).
2. While the Webb and Achieve approaches measure alignment along similar dimensions, Achieve is much more qualitative and rejects the consensus approach of the Webb methodology.
3. The Achieve approach also considers the role of test blueprints in translating standards into assessments, whereas the Webb and SEC approaches ignore blueprints.
4. The SEC approach is the only approach that can be used to measure alignment of other documents (e.g., curriculum materials, survey reports of teachers' instruction) with standards and assessments.
5. The Webb approach offers thresholds for determining the adequacy of alignment on each of its dimensions, while the SEC does not, but instead offers techniques for hypothesis testing of alignment indices.

Forte (2017) concluded that the choice of alignment method “should depend most on what information the state would find most helpful in improving the quality of its systems of standards and assessments in support of instruction” (p. 11).

How Well Aligned Are State Tests With State Standards?³

The main role of alignment methods is to investigate the alignment of tests with standards. Under NCLB and ESSA, states must submit evidence for peer review that their assessments are meaningfully aligned with their state standards. States have typically used the Webb procedure or a variant to conduct these studies (Martone & Sireci, 2009). Evidently, peer reviewers have been satisfied with the alignment evidence that has been produced, because these assessments continued to be developed and administered for accountability. Early Webb alignment studies (e.g., Webb, 2002) identified some common alignment issues that assessments had: failing to cover the full breadth of the standards, failing to cover the objectives under a standard, and failing to reach the adequate depth of knowledge of the standards. In that study, only one of the eight tests studied met all four alignment criteria. More recent studies (e.g., Webb, 2005, 2010) have generally found more acceptable levels of alignment of state tests with state standards and weaker alignment of other tests (e.g., ACT/SAT) with state standards (e.g., NORC, 2015; Roeber et al., 2018).

Test-to-standards alignment studies have also been conducted using the SEC. The most comprehensive of these was a 2011 analysis (Polikoff et al., 2011) that compared 138 pairs of state standards and assessments in mathematics, science, and ELA. That analysis found that assessments (a) routinely overemphasized some content in the standards and underemphasized other content, (b) assessed content at lower levels of cognitive demand than was called for in the standards, and (c) contained sometimes 20%–25% of content that was not in the corresponding grade-level standards. Furthermore, certain content areas of state standards were routinely not tested year after year.

A more recent analysis also investigated the alignment of state assessments, this time in the Common Core era (Doorey & Polikoff, 2016). This study did not consider alignment the way the Webb or SEC approaches do. Rather, it examined alignment in terms of key instructional shifts called for by the Common Core standards. It found that the new generation of assessments—specifically PARCC and Smarter Balanced—did outperform a “best-in-class” state assessment (MCAS) in terms of many of these key shifts. For instance, the new assessments were better at emphasizing close reading, writing to texts, research and inquiry skills, and matching the cognitive complexity of the standards than were MCAS. However, the ACT Aspire—another test intended to align to Common Core standards—scored lower than PARCC and Smarter Balanced (and, indeed, the MCAS) on several dimensions. The PARCC and Smarter Balanced tests did fall short of aligning to standards in several ways. For instance, they failed to assess speaking and listening standards, and they did not match the complexity of the standards in mathematics. Separately, there is also the recent special case of states using college entrance examinations (ACT, SAT) as their summative assessments for high school. These assessments were not constructed for—and are plainly not aligned

to—state high school standards, though they may offer other benefits such as boosting college awareness and enrollment among groups that may be less likely to otherwise enroll in college (Lorié, 2020).

Together, these studies do suggest that state tests have sometimes struggled in certain key areas with regard to alignment to standards. In particular, they have struggled to assess the full cognitive complexity of the standards (falling short of capturing the most cognitively complex content). This is perhaps not surprising because the most cognitively complex content is often difficult to assess unless there is an extended period or some degree of student research and writing. State tests have also sometimes failed to assess key chunks of standards content (for instance, NCLB-era standards overemphasized reading skills and underemphasized writing skills relative to the standards), perhaps undermining the content messages of the standards somewhat. The differences in findings across studies also suggest that different alignment approaches (which are conducted with different groups of raters) may lead to different answers as to how well aligned assessments are with standards. There is no one “best” alignment method, nor is it even clear how one would compare results from different alignment methods to reach broader implications, which makes it difficult to make definitive statements about overall test-to-standards alignment.

KORETZIAN INFLATION THEORY AND THE VALIDATION OF TEST-BASED INFERENCES IN ACCOUNTABILITY SYSTEMS

Alignment also plays a role in the implications and evaluation of test-based inflation, where systematic misalignment creates opportunities for students, teachers, and systems to attend to construct-irrelevant features of tests. We use the term Koretzian inflation theory to credit the empirical and theoretical work of Daniel Koretz and his coauthors toward the validation of test-based inferences under high-stakes conditions (Holcombe et al., 2013; Koretz, 2008, 2015; Koretz & Barron, 1998; Koretz & Hamilton, 2006; Koretz et al., 2001, 2016). Koretzian inflation theory is a corollary of Campbell’s law (Campbell, 1979) as it applies to high-stakes test-based educational inferences. Campbell’s law states, “the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor” (p. 49). Koretz and Hamilton (2006) described inflated test-based inferences as follows: “Test scores can become inflated—that is, can be higher than proficiency in the measured domain warrants . . . particularly severe[ly] when the consequences for scores are substantial” (p. 542).

Empirical studies of Koretzian inflation theory ask a counterfactual question of any high-stakes inference drawn from test scores: How would the inference differ had we used a lower stakes test of the intended domain? The theory predicts that the inference drawn from the high-stakes (focal) test is inflated and biased over the inference drawn from a well-aligned low-stakes (audit) test. The canonical case is that of cohort-to-

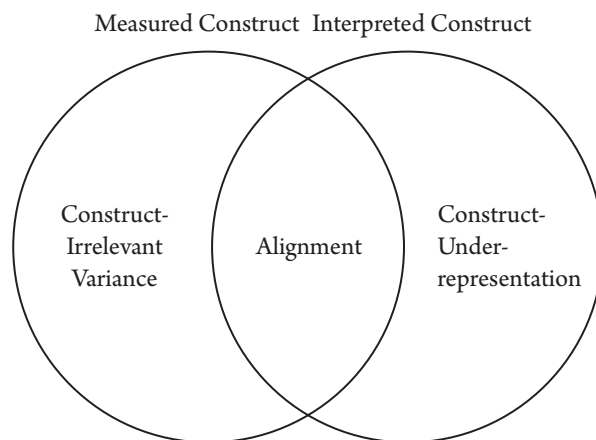
cohort measures of average progress such as those from the NCLB-era New York State test, which from 2005 to 2009 showed an increase of over half a standard deviation in Grade 8 Mathematics. In contrast, the cohort-to-cohort NAEP progress in New York from 2005 to 2009 was less than 0.1 standard deviations (Koretz, 2017).

Without a clearer specification of the intended inference and measurement, discrepant trends are not necessarily evidence of inflation. Ho (2007) summarized state–NAEP discrepancies from 26 states from 2003 to 2005 and found that state test score trends from 2003 to 2005 were approximately four times the magnitude of NAEP score trends (0.12 to 0.03 standard deviation units over 2 years). Jacob (2007) looked at four states from 1992 to 2003 and found focal test score trends of 0.09 and 0.06 standard deviations per year in math and reading, respectively, compared to 0.05 and 0.02 standard deviations per year on NAEP.

How can we explain discrepant trends? Any differences in differences in sampling, administration conditions, or behaviors may result in a trend discrepancy. For example, this pattern of trend discrepancies might arise if lower proficiency students began to opt out of the high-stakes test but not NAEP. Student motivation may increase disproportionately for the state test over NAEP. Teachers may increase the alignment of the curriculum to tested standards for the state test over NAEP. Teachers may begin to drill tested standards for the state test over NAEP. Or students or teachers may begin to cheat on the state test but not on NAEP. In all of these cases, it is differences in differences, or the interaction effect, not just differences between tests, that account for trend discrepancies.

In the case where trends differ because of what is measured, which one is correct? Koretzian inflation theory (Koretz, 2008; Koretz & Hamilton, 2006; Koretz et al., 2001) is an applied validation framework that emphasizes a sampling principle of testing and, because it is most commonly applied to measures of progress, a distinct temporal dimension. We present a simplified version of the framework in Figure 16.2, where a test and a target of inference are intersecting circles of a Venn diagram full of measurable and inferable elements of performance. This is a common depiction where the tested (but not inferred) section represents construct-irrelevant variance and the inferred (but not tested) section represents construct underrepresentation. If what is tested is a random sample of the population of elements and the construct-irrelevant variance is random, then we appear to be assured of valid inferences.

Koretzian inflation theory adds a temporal dimension to this representation under high-stakes conditions. Applying Campbell's law and a sampling principle of testing, it asserts that even an initially random sample of tested and targeted elements can become the focus of disproportionate weight or outright corruption, if that sample becomes predictable over time. Students and teachers may notice an initially representative subsample and begin to weight it disproportionately over desired inference weights. This may lead to inflation if, for example, such elements of performance are efficient psychometrically and end up being reused or because such elements form the basis of an equating item set that ensures comparability over time.

**FIGURE 16.2**

Illustrating Mechanisms for Inflation When Test Developers Sample Predictably From Content Domains

The theory defines *reallocation* as a process of increasing weights on tested content disproportionately over their inference weights over time. Reallocation also occurs when teachers and students decrease weights on untested elements that are nonetheless part of the target inference. This is commonly called *narrowing the curriculum* or *teaching to the test*, and we discussed evidence of its prevalence in previous sections. Koretzian inflation theory suggests that no test can simply be “worth teaching to.” As long as a test is a predictable sample of desired elements, stakes will lead to invalid interpretations of progress. Reallocation degrades inferences at the right side of Figure 16.2, increasing emphasis on the intersection (what is measured) at the expense of important but underrepresented content.

Similarly, initially ignorable construct-irrelevant variance can also inflate progress over time under high-stakes conditions. The theory defines *coaching* as the disproportionate weighting of tested elements that are not part of the target inference. This may include elements like “process of elimination,” where students are coached to eliminate obviously incorrect options in a multiple-choice item and, if necessary, guess among the remaining possibilities. Coaching works at the left side of Figure 16.2 by increasing emphasis on construct-irrelevant but predictable features. To the extent that eliminating possibly incorrect answers is not part of the target of inference for, say, a reading test score, Koretz (2017) argued that coaching is at best a waste of instructional time and at worst a source of bias and inflation in test scores.

Holcombe et al. (2013) added multiple levels to the simple heuristic figure sketched in Figure 16.2. The opportunities for score inflation are numerous and appear at every stage of the standard development of tests, including selecting the domain, selecting the elements of the domain, selecting standards for testing, and selecting item representations from tested standards. Each stage introduces possible construct underrepresentation and construct-irrelevant variance. The key insight is that this potential for inflation rises

with higher stakes on results, incentivizing educators to employ coaching and gaming and resulting in inflated indicators of progress.

Koretzian inflation theory also motivates a number of tools and solutions. If high stakes on individual measures are necessary, the theory clearly motivates the use of a low-stakes audit test. NAEP has served that *de facto* role for states and urban districts since the reporting of state and district scores in the 1990s. Although state and district leaders are held publicly accountable to NAEP progress via transparent reporting, the stakes are not as high or as direct as they are for school leaders and teachers. Numerous district tests may also serve lower stakes purposes and can serve as audit measures, although they may not be uniform and comparable at the state level.

Koretz and Beguin (2010) proposed self-monitoring assessments that embed two types of audit items into operational tests. A content audit item would assess a desired target of inference that may be underrepresented in operational items. A style audit item would measure existing content in an unpredictable way to protect against possible inflation from coaching. Koretz et al. (2016) concluded that the approach had theoretical promise but faced substantial practical constraints given the already limited space for items on operational tests. There are also potentially disproportionate costs when developing items from underrepresented content domains that may be underrepresented because they are costly to measure.

Neal (2013, 2018) proposed an alternative solution whereby two separate tests or test sections would achieve separate purposes, the first to measure progress and the second to estimate the causal effects of teachers and schools on test scores. The first test would be largely stable to measure change over time but have no direct stakes on teachers and schools. The second test could change substantially from year to year to sample broadly from the domain with no constraints on equating items, balance across content areas, or consistent item formats. Such a test would render reallocation and coaching difficult or impossible and theoretically encourage educators to focus on the curriculum standards rather than construct-irrelevant performance elements. Because models that estimate teacher and school effects regress current-year scores on past-year scores, they do not need to be equated to the same scale.

Koretz (2017) offered additional policy solutions, the most straightforward of which is to reduce stakes or dilute them through the use of a multiple-measure indicator system. Other solutions include setting reasonable and achievable targets and providing supports that are commensurate with stakes in a reciprocal accountability system (Elmore, 2004).

ACCOUNTABILITY TESTING FOR SPECIAL POPULATIONS

The logic of test-based accountability suggests principles of inclusion and comparability for all subpopulations. Following Figure 16.1, if students from special populations

are excluded or not disaggregated, it sends a signal that they are not a priority. If student scores are not comparable or counted, then the incentives to teach them become uncertain or diminished. In this section, we briefly discuss the inclusion of students with disabilities (SWDs) and English learners (ELs, sometimes also called limited English proficient, English language learners, or emergent bilinguals). Our review is necessarily brief and focuses primarily on issues directly related to test-based accountability for students in these groups. For additional consideration of these groups, see Zwick and Rodriguez and Thurlow in this volume.

NCLB's accountability provisions brought substantial focus on the measurement of all SWDs and ELs. From the earliest days of NCLB, scholars called attention to the salient accountability issues for these groups. Issues related to SWDs were prominent under NCLB, with NCLB policy specifically identifying two groups: (a) up to 1% of the tested population—those with the most severe cognitive disabilities—to be measured against alternate achievement standards; and (b) up to 2% of students with less severe disabilities but who would be unlikely to meet grade-level standards to be measured against modified achievement standards (Elledge et al., 2009). And NCLB is credited with bringing about substantial change and standardization in the processes used to assess EL students within and across states.

McLaughlin and Thurlow (2003) described some of the major substantive and technical issues related to educational accountability for students with disabilities. The authors first noted the inherent tensions of standards-based accountability with key provisions of special education law, such as the guarantees of the Individuals with Disabilities Education Act. The heart of federal law for SWDs is the individualized education program, which is not generally standards based and focuses on each child's individualized performance, goals, services, and participation in the general education classroom. In contrast, NCLB's accountability provisions required participation in the state's standards-based assessment for all but a small percentage of students with the most severe disabilities. The very nature of content standards themselves seems to run counter to the individualized nature of special education services, and special education scholars have questioned the appropriateness of standards for some SWDs. For students given alternate assessments based on the nature of their disabilities, McLaughlin and Thurlow discussed the challenges of aggregating these results with the results of the general population, including setting appropriate and attainable goals for these students and issues of confidentiality based on small group sizes. They also discussed test modifications commonly offered to SWDs and the potential for these modifications to affect the validity of inferences made on the basis of student performance. And for SWDs who are excluded for some or all of the general education curriculum, there are concerns about students' opportunity to learn and the validity of the assessment results for inferences about the general education curriculum.

Similarly, Abedi and Dietel (2004) summarized some of the major substantive and technical issues related to educational accountability for ELs. First, EL students are typically among the lowest-performing groups on content-area assessments, so the uniform

accountability targets of NCLB were seen as especially—and perhaps unrealistically—ambitious for this group. Second, EL students are especially sensitive to the language demands of accountability tests, and Abedi and Dietel made clear that these tests measure both students' content knowledge and their language ability. Third, EL students transition out of the category over time as their English language proficiency improves. Because new EL students often have low scores and high-scoring EL students transition out of the category, the average achievement of the EL category can often be low despite academic growth of EL students. If accountability systems incentivize average proficiency for EL students without attention to accurate reclassification, this creates an incentive to retain English learners in the EL category even after they may benefit from these services.

Other issues raised by Abedi and Dietel (2004) include the variation in approaches and assessments used to identify EL students, the diversity among EL subgroups (e.g., Chinese-speaking ELs versus Spanish-speaking ELs), and the tautological relationship between English language proficiency and EL classification (i.e., if all students were proficient in English, as was NCLB's goal, there would be no EL students by definition).

In the prior edition of this volume, Koretz and Hamilton (2006) discussed these and other issues relating to SWDs and ELs in the context of assessment for accountability. For instance, they discussed the heterogeneity of SWD and EL populations and noted the implications for group size (e.g., many EL or SWD groups are too small to be numerically significant for reporting). And they considered specific disability categories and the implications of students' disability types for validity issues related to construct relevance (e.g., a dyslexic student's challenges in reading text in the context of a mathematics assessment). The primary focus of their treatment of SWD and EL issues is their more extensive discussion of accommodations and modifications and their implications for validity. Changes in accommodation and modification policies brought on by NCLB's inclusion mandates had important implications for the kinds of students included in state assessment and accountability systems. While these changes brought needed attention to the academic performance of students with disabilities and raised expectations for many of these students, scholars have questioned the appropriateness of grade-level standards for many students with disabilities.

There are several important innovations pertaining to accountability assessment for ELs and SWDs under ESSA policy. These innovations broadly attempt to deal with some of the unintended consequences and issues that emerged under NCLB-era regulations. For instance, ESSA regulations do away with the 2% group for modified standards, while maintaining the 1% group with the most severe disabilities who can be tested against alternate standards. This addresses concerns that too many students were being excluded from accountability against regular grade-level standards with the 2% cap. Addressing the tension mentioned above between individualized education programs and inclusion, ESSA regulations specifically require states to have guidelines for the inclusion or exclusion of students with disabilities in the alternate assessment

system. In principle, this could again improve the appropriateness of assessments offered to students with disabilities. ESSA law also emphasizes the importance of universal design for learning in the test development process, which it defines as follows:

The term “universal design for learning” means a scientifically valid framework for guiding educational practice that—(A) provides flexibility in the ways information is presented, in the ways students respond or demonstrate knowledge and skills, and in the ways students are engaged; and (B) reduces barriers in instruction, provides appropriate accommodations, supports, and challenges, and maintains high achievement expectations for all students, including students with disabilities and students who are limited English proficient. Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). [congress.gov/114/plaws/publ95/PLAW-114publ95.pdf](https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf)

This is intended to reduce concerns about construct irrelevance, some of which were highlighted by Koretz and Hamilton (2006). For more on the implications of the universal design for learning in the context of accountability assessments, see Rose et al. (2018). For more discussion of alternate assessments and the English Learner population, also see Rodriguez and Thurlow in this volume.

PERSPECTIVES ON EQUITY IN ACCOUNTABILITY TESTING

Educational policies like ESSA have stated aims and mechanisms to provide low-income students with improved educational opportunities. This reflects one definition of equity, the disproportionate allocation of resources to those with greatest need. The intended role of educational tests in these policies is to measure students’ academic needs so that educational systems can similarly provide low-scoring students with disproportionate support. As we have reviewed in earlier sections in this chapter, there are unintended consequences when these systems place unrealistic expectations on teachers and school leaders to raise test scores for low-scoring students. In these cases, attempts to improve equity for low-scoring students may be inequitable for the teachers and leaders who support these students.

Tests may also fail to improve equity when they are disproportionately burdensome to the lowest-scoring students who take these tests. This can occur in the act of testing, when low-scoring students encounter assessment tasks and prompts that are not as engaging or affirming to them as they are to higher scoring students, or when they receive test scores, when low-scoring students may disproportionately interpret their low test scores as evidence of their irredeemable deficits rather than areas for potential growth. This section reviews selected perspectives on equity in accountability testing.

Equity Monitoring Systems

In the 2019 NASEM report, *Monitoring Educational Equity*, the authors argued that relying primarily on test scores to measure school and student performance can

undermine educational equity by neglecting the inputs and context for learning. A more comprehensive system could reveal inequality earlier in the life span, set realistic expectations for student growth, and provide national benchmarks to spur investment and improvement. Such a system would consider factors like school funding levels, curriculum quality, teacher qualifications, class sizes, student health and well-being, extra-curricular opportunities, and other resources that impact learning.

The report recommended that the indicators track inequality in opportunities at the school and district level, as well as for various student demographic subgroups. This would provide insights into how policies differentially impact groups like students of color, low-income students, ELs, and SWDs. Monitoring progress in closing equity gaps could then help target resources and support. Comparisons of progress across the country could inspire learning from proven practices.

The report emphasized that developing a national monitoring system will require considerable input to secure buy-in before full implementation. Using NAEP as an example, the authors proposed a board like NAGB that oversees and reports on a system of comparable equity indicators across states and over time, inspiring improvement through public accountability. They argued that indicators must be selected carefully and with appropriately developed and publicly reported validity evidence. Ongoing evaluation would also be needed to ensure the system meets its goals of revealing inequities and spurring corrective action over time.

Returning to Figure 16.1, adopting an “equity monitoring” approach as proposed by these authors holds promise to both reveal and reduce the unintended negative consequences listed on the right-hand side of the framework. For example, highlighting inputs rather than outputs can help to set realistic expectations for growth. To the extent that there are stakes through public accountability, these stakes would be more distributed across additional indicators beyond test scores, reducing the likelihood of inflation. Unintended consequences of public reporting of indicators, like segregation, would be monitored directly. And systematic attention to disaggregating indicator data by sociodemographic categories, particularly for indicators related to educational inputs, would help to avoid deficit frameworks by highlighting inequality in educational access and opportunities (Quinn & Desruisseaux, 2022).

Culturally Sustaining Assessment and Test-Based Accountability

If the NASEM report is an effort to “measure equity” by documenting disparities in a broader range of school- and student-related indicators, there is also increased effort to “measure equitably.” This effort involves measuring outputs like achievement with particular investment in understanding the learning experiences of historically marginalized students and designing assessments that engage them. This builds on theories of “culturally sustaining pedagogies” (Ladson-Billings, 1995; Paris, 2012), which are educational practices that not only acknowledge but also build on the cultural

knowledge and experiences students bring to the classroom. In classroom contexts, culturally sustaining assessments are activities designed to value, engage, and empower students (Lyons et al., 2021; Randall, 2021). (In this volume, see Ercikan & Flores; Lane & Marion; Rodriguez & Thurlow; and Zwick for discussion of culturally responsive assessments.)

Figure 16.1 distinguishes between direct score-based actions and indirect non-score-based influencing actions. For culturally sustaining assessments to function well in accountability systems, they must have positive direct and indirect actions. Scores from culturally sustaining assessments would need to accurately identify both schools and students who need support. And the practice of culturally sustaining assessment would need to improve agency and engagement among students and educators. An unintended negative consequence would be if accountability pressures on educators led them to customize tests under the banner of “culturally sustaining assessments” that inflated scores for students who needed support, thus decreasing the likelihood that they would receive the support they need.

Dee and Penner (2017) provided evidence that an ethnic studies course designed with principles from culturally sustaining pedagogy has positive effects on student attendance and grade point average in other courses. Using a regression discontinuity approach, they estimated that students near the assignment cutoff (an eighth-grade grade point average of 2.0) would have a 1.4-point higher ninth-grade grade point average had they been assigned to take the course, compared to the counterfactual of not being assigned. Bonilla et al. (2021) showed that there are still detectable causal effects in terms of high school attendance, high school graduation, and postsecondary matriculation. However, the ethnic studies course was not designed with culturally sustaining assessment for accountability as a goal, nor were the properties of scores or grades from this course a target of analysis.

Theoretical foundations for culturally sustaining testing exist in sociocognitive models for learning and assessment (Bennett, 2023; Lyons et al., 2021; Mislevy, 2018; Moss et al., 2008; NASEM, 2018; Shepard, 2021). Sociocognitive models incorporate the past experiences of students and the current context of learning and assessment, including students’ relationships with other students and the teacher in the classroom. These models typically result in inferences that are richer and context referenced and thus are more limited in claims about generalizability. The sociocognitive perspective on conventional large-scale tests is that these are also context referenced to isolated individual completion of a particular subset of decontextualized cognitive tasks. Scores on these tests are then similarly limited in generalizability to other performances on tasks that also require isolated individual completion of decontextualized cognitive tasks.

Thus, although there is optimism around culturally sustaining assessment that serves students and produces low-stakes scores for classroom inferences and scholarly research (Bennett, 2023; Mislevy, 2018; Shepard, 2021), there is skepticism about

the role that scores from culturally sustaining tests might play in accountability systems (Lyons et al., 2021). Stakes would have to remain low to avoid incentives to change assessment tasks to inflate scores. Traditional psychometric frameworks that require strict comparability of score interpretations across contexts would have to evolve. Costs for test development, deployment, evaluation, and scoring would have to be feasible. And building scholarly, political, and community consensus about content standards and test blueprints for culturally sustaining tests would take time and work.

As one example of such a process, the state of Hawaii has allowed Native Hawaiians to design a test and test-based accountability system for students in the Kaiapuni Educational Program, a Hawaiian language immersion program. The Kaiapuni Assessment of Educational Outcomes (KĀ'EO) follows ESSA guidelines, including the requirement for mandatory public reporting in terms of KĀ'EO standards in mathematics, Hawaiian language arts, and science. According to the KĀ'EO technical report (Hawaii State Department of Education, 2018), the development team used the achievement-level descriptors for the English language state test, the Smarter Balanced test, as a foundation for the development of Hawaiian language achievement-level descriptors. These, in turn, guided item writing and standard setting in Hawaiian. Shultz and Englert (2021) described this as “one example of how the test developers and community stakeholders successfully balanced the tension between maintaining the technical requirements of a state assessment and serving the needs of the community as defined by the community” (p. 5).

CONCLUSION

In a first draft of this chapter on test-based accountability, written before the onset of the COVID-19 pandemic, our opening sentence was, “Over the past several decades, the arc of K–12 educational assessments has bent steadily toward accountability.” At the time, although ESSA had granted some flexibility from one-size-fits-all accountability, the statement generally seemed defensible. Years later, the sentence seems laughably dated, a reminder of the pendulum swings that are common between extremes of federal and local control and the imprecision of extrapolating from past trends.

Although predicting future trends in test-based accountability is not our goal, we believe the principles outlined in the beginning of this chapter and revisited throughout should guide future validation of test score use in educational accountability systems. Once educational test scores exist in complex systems, actors use them for a predictable range of purposes that extend well beyond simple descriptions of what individual students know and can do. Validation, and the corresponding anticipation of unintended negative consequences, requires specification of actors, score-based actions, score types, and indirect actions in social and political systems (Figure 16.1).

ACKNOWLEDGMENTS

The authors thank the coeditors, Linda Cook and Mary Pitoniak, the designated reviewers of this chapter, Laura Hamilton and Joan Herman, and the Editorial Advisory Board reviewer, Michael Rodriguez, for their constructive feedback and insights. We also thank Lily An and additional members of the Harvard Graduate School of Education's Educational Measurement Lab for their suggestions.

REFERENCES

- Abdulkadiroglu, A., Pathak, P., Schellenberg, J., & Walters, C. (2017). *Do parents value school effectiveness?* (Working Paper No. 23912). National Bureau of Economic Research.
- Abedi, J., & Dietel, R. (2004). Challenges in the No Child Left Behind Act for English-language learners. *Phi Delta Kappan*, 85(10), 782–785.
- Achieve. (2004). *Do graduation tests measure up? A closer look at state high school exit exams*.
- Ahn, T., & Vigdor, J. (2014). *The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina* (NBER Working Paper No. 20511). National Bureau of Economic Research.
- Airasian, P. W. (1987). State mandated testing and educational reform: Context and consequences. *American Journal of Education*, 95(3), 393–412.
- Alexander, L., James, H. T., & Glaser, R. (1987). *The Nation's Report Card: Improving the assessment of student achievement*. National Academy of Education.
- Allensworth, E., & Easton, J. (2005). *The on-track indicator as a predictor of high school graduation*. The University of Chicago Consortium on School Research.
- Alliance for Excellent Education. (2018). *N-size in ESSA state plans*.
- Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve*. Education Trust.
- American Educational Research Association. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher*, 44(8), 448–452.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Institutes for Research. (2017). *Examining ESSA plans through the lens of research and practice: Reflections on state ESSA plans*.
- American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*.
- Ansari, A., & Pianta, R. C. (2019). School absenteeism in the first decade of education and outcomes in adolescence. *Journal of School Psychology*, 76, 48–61.
- Armor, D. J., & Peiser, B. M. (1998). Interdistrict choice in Massachusetts. In P. E. Peterson & B. C. Hassel (Eds.), *Learning from school choice* (pp. 157–186). Brookings Institution Press.

- Atchison, D. (2020). The impact of Priority School designation under ESEA flexibility in New York State. *Journal for Research on Educational Effectiveness*, 13(1), 121–146.
- Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (NBER Working Paper No. 20657). National Bureau of Economic Research.
- Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research.
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute.
- Baker, O., & Lang, K. (2013). *The effect of high school exit exams on graduation, employment, wages, and incarceration* (NBER Working Paper 19182). National Bureau for Economic Research.
- Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007*. National Center for Education Statistics.
- Bayer, P., Ferreira, F., & McMillan, R. (2007). *A unified framework for measuring preferences for schools and neighborhoods* (Working Paper No. 13236). National Bureau of Economic Research.
- Bellwether Education Partners. (2017). *An independent review of ESSA state plans*.
- Bennett, R. E. (2023). Toward a theory of socioculturally responsive assessment. *Educational Assessment*, 28(2), 83–104.
- Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, 52(5) 861–893.
- Betebenner, D. W. (2008). *A primer on student growth percentiles*. National Center for the Improvement of Educational Assessment.
- Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51.
- Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2023). SGP: Student growth percentiles & percentile growth trajectories (Version 2.1-0.0) [R package]. <https://cran.r-project.org/web/packages/SGP/SGP.pdf>
- Billingham, C. M., & Hunt, M. (2016). School racial composition and parental choice new evidence on the preferences of White parents in the United States. *Sociology of Education*, 89(2), 99–117.
- Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, 114(2), 577–599.
- Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, 39(1), 146–170.
- Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M., & Springer, M. (2024). Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment. *Journal of Political Economy: Microeconomics*. Advance online publication.

- Blinder, A. (2015, April 1). Atlanta educators convicted in school cheating scandal. *The New York Times*. <https://www.nytimes.com/2015/04/02/us/verdict-reached-in-atlanta-school-testing-trial.html>
- Bloom, B. S. (1968). Mastery learning. *Evaluation Comment*, 1(2), 1–11.
- Bloom, B. S. (1973). *Cross-national study of educational attainment: Stage 1 of the I.E.A. investigation in six subject areas*. U.S. Department of Health, Education, & Welfare.
- Bonilla, S., & Dee, T. (2020). The effects of school reform under NCLB waivers: Evidence from focus schools in Kentucky. *Education Finance and Policy*, 15(1), 75–103.
- Bonilla, S., Dee, T. S., & Penner, E. K. (2021). Ethnic studies increases longer-run academic engagement and attainment. *Proceedings of the National Academy of Sciences*, 118(37), e2026386118. <https://www.pnas.org/doi/10.1073/pnas.2026386118>
- Booher-Jennings, J. (2005). Below the bubble: “Educational triage” and the Texas accountability system. *American Educational Research Journal*, 42, 231–268.
- Boser, U., & Rosenthal, L. (2012). *Do schools challenge our students? What student surveys tell us about the state of education in the United States*. Center for American Progress.
- Buckley, K. H. (2015). Separating the signal from the noise: An examination of student and teacher scores based on student learning objectives (SLOs) in one state. [Doctoral dissertation, Harvard University]. <https://dash.harvard.edu/bitstream/handle/1/16461041/BUCKLEY-DISSERTATION-2015.pdf?sequence=1>
- Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education*, 85(2), 147–162.
- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.
- Cannata, M. A., Smith, T. M., & Haynes, K. T. (2017). Integrating academic press and support by increasing student ownership and responsibility. *AERA Open*, 3(3), 1–13.
- Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305–331.
- Castellano, K. E., & Ho, A. D. (2013a). *A practitioner’s guide to growth models*. Council of Chief State School Officers.
- Castellano, K. E., & Ho, A. D. (2013b). Contrasting OLS and quantile regression approaches to student “growth” percentiles. *Journal of Educational and Behavioral Statistics*, 38(2), 190–215.
- Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics*, 40(1), 35–68.
- Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for high school graduation. *American Journal of Education*, 98(1), 1–34.
- Center on Education Policy. (2008). *State high school exit exams: A move toward end-of-course exams*.
- Center on Education Policy. (2009). *State high school exit exams: Trends in test programs, alternate pathways, and pass rates*.

- Center on Education Policy. (2012). *State high school exit exams: A policy in transition*.
- Chakrabarti, R. (2014). Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics*, 110, 124–146.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9), 2633–2679.
- Chetty, R., Friedman, J., & Rockoff, J. (2014b). Discussion of the American Statistical Association's Statement (2014) on using value-added models for educational assessment. *Statistics and Public Policy*, 1(1), 111–113.
- Chudowsky, N., Kober, N., Gayler, K. S., & Hamilton, M. (2002). *State high school exit exams: A baseline report*. Center on Education Policy.
- Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, 12(3), 311–329.
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378–387.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. U.S. Government Printing Office.
- Corcoran, S., Jennings, J., & Beveridge, A. (2011, March 3–5). *Teacher effectiveness on high- and low-stakes tests* [Paper presentation]. Society for Research on Educational Effectiveness Spring Conference, Washington, DC, United States.
- Council of Chief State School Officers. (1984). *Education evaluation and assessment in the United States*.
- Cronbach, L. J. (2004). An interview with Lee J. Cronbach. In L. V. Jones & I. Olkin (Eds.), *The nation's report card: evolution and perspectives* (pp. 139–153). Phi Delta Kappa Educational Foundation.
- Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. (2005). *The impact of the No Child Left Behind Act on student achievement and growth*. Northwest Evaluation Association.
- D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Measurement*, 12(1), 1–22.
- Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, 22(86), 1–34.
- Data Quality Campaign. (2019). *Show me the data [Report]*. <https://dataqualitycampaign.org/wp-content/uploads/2019/04/DQC-Show-Me-the-Data-04042019.pdf>.
- Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher*, 44(6), 347–358.

- Dee, T. S. (2003). The “first wave” of accountability. In P. Peterson & M. West (Eds.), *No Child Left Behind? The politics and practice of accountability* (pp. 1–20). Brookings Institution.
- Dee, T., & Dizon-Ross, E. (2017). *School performance, accountability and waiver reforms: Evidence from Louisiana* (NBER Working Paper No. 23463). National Bureau of Economic Research.
- Dee, T. S., & Jacob, B. A. (2006). *Do high school exit exams influence educational attainment or labor market performance?* (NBER Working Paper 12199). National Bureau of Economic Research.
- Dee, T. S., & Jacob, B. A. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, 30, 418–446.
- Dee, T. S., Jacob, B. A., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, 35(2), 252–279.
- Dee, T. S., & Penner, E. K. (2017). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. *American Educational Research Journal*, 54(1), 127–166.
- Delaware Department of Education. (2010). *For the 2009–2010 school year: State accountability in Delaware*. http://www.doe.k12.de.us/aab/accountability/Accountability_Files/School_Acct_2009-2010.pdf
- Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics*, 98(5), 848–862.
- Domaleski, C. (2011). *State end-of-course testing programs: A policy brief*. Council of Chief State School Officers.
- Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Center for American Progress.
- Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Thomas B. Fordham Institute.
- Education Commission of the States. (2014). *Rating states, grading schools: What parents and experts say states should consider to make school accountability systems meaningful*.
- Education Commission of the States. (2018). *50-state comparison: States’ school accountability systems*.
- Edwards, N. R., & Mindrila, D. L. (2019). Improving graduation rates: Legitimate strategies and gaming practices. *Education Policy Analysis Archives*, 27(41), 1–28.
- Elledge, A., Le Floch, K. C., Taylor, J., & Anderson, L. (2009). *State and local implementation of the No Child Left Behind Act Volume V—Implementation of the 1 percent rule and 2 percent interim policy options*. U.S. Department of Education.
- Elliott, S. N., & Bartlett, B. (2016). Opportunity to learn. In *Oxford handbooks online scholarly research reviews* (pp. 1–11). Oxford University Press.

- Elmore, R. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In S. Fuhrman & R. Elmore (Eds), *Redesigning accountability systems for education* (pp. 274 – 296). Teachers College Press.
- English, D. (2017). *Proposed state accountability systems under the Every Student Succeeds Act: A summary of fall 2017 submissions*. American Institutes for Research.
- Eren, O., Depew, N., & Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, 153, 9–31.
- Eren, O., Lovenheim, M. F., & Mocan, N. H. (2018). *The effect of grade retention on adult crime: Evidence from a test-based promotion policy* (NBER Working Paper 25384). National Bureau for Economic Research.
- Figlio, D. N., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. J. Machin, & L. Woessmann (Eds.), *Handbooks in economics: Economics of education* (Vol. 3, pp. 383–421). Elsevier.
- Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review*, 94(3), 591–604.
- Figlio, D., & Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90(1–2), 239–255.
- Figlio, D. N., Rouse, C. E., & Schlosser, A. (2009). *Leaving No Child Behind: Two paths to school accountability* [Working paper]. The Urban Institute.
- Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems*. edCount.
- Fuchsman, D., Sass, T. R., & Zamarro, G. (2020). *Testing, teacher turnover and the distribution of teachers across grades and schools* (Annenberg Ed Working Paper No. 20–200). Brown University.
- Gaddis, S. M., & Lauen, D. L. (2014). School accountability and the Black–White test score gap. *Social Science Research*, 44, 15–31.
- Gewertz, C. (2020). Which states require an exam to graduate? *Education Week*. <https://www.edweek.org/ew/section/multimedia/states-require-exam-to-graduate.html>
- Ginsburg, A. L., Noell, J., & Plisko, V. W. (1988). Lessons from the wall chart. *Educational Evaluation and Policy Analysis*, 10(1), 1–12.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18(8), 519–521.
- Glazerman, S., & Dotter, D. (2017). Market signals: Evidence on the determinants and consequences of school choice from a citywide lottery. *Educational Evaluation and Policy Analysis*, 39(4), 593–619.
- Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added*. Brown Center on Education Policy at Brookings.
- Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, 36, 216–228.
- Greene, J. P., & Winters, M. A. (2007). Revisiting grade retention: An evaluation of Florida's test-based promotion policy. *Education Finance and Policy*, 2(4), 319–340.

- Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas*. National Education Goals Panel.
- Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis*, 36(4), 417–436.
- Grodsky, E., Warren, J. R., & Kalogrides, D. (2009). State high school exit examinations and NAEP long-term trends in reading and mathematics, 1971–2004. *Educational Policy*, 23(4), 589–614.
- Hackmann, D. G., Malin, J. R., & Bragg, D. D. (2019). An analysis of college and career readiness emphasis in ESSA state accountability plans. *Education Policy Analysis Archives*, 27(160), 1–28.
- Haderlein, S. A. K. (2022). How do parents evaluate and select schools? Evidence from a survey experiment. *American Educational Research Journal*, 59(2), 381–414.
- Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Interdisciplinary research and Perspectives*, 11(1–2), 1–18.
- Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education*, 104(2), 1–34.
- Haertel, E., & Ho, A. (2016). Fairness using derived scores. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 233–254). Routledge.
- Hamilton, L. S. (2012). Measuring teaching quality using student achievement tests: Lessons from educators' responses to No Child Left Behind. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 49–75). Teachers College Press.
- Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
- Harris, D., & Herrington, C. (2004). *Accountability and the achievement gap: Evidence from NAEP* [Unpublished manuscript]. Department of Educational Leadership and Policy Studies, Florida State University.
- Harris, D. N., & Larsen, M. F. (2017). *Demand, information, and the market for schooling: Evidence on revealed preferences from post-Katrina New Orleans*. Education Research Alliance for New Orleans.
- Harris, D. N., Liu, L., Barrett, N., & Li, R. (2020). *Is the rise in high school graduation rates real? High-stakes school accountability and strategic behavior*. Brookings Institution.
- Hastings, J. S., Kane, T. J., & Staiger, D. O. (2009). *Heterogeneous preferences and the efficacy of public school choice*. Yale University.
- Hawaii State Department of Education. (2018). *KĀ'EO 2018 technical manual for grades 3-4 HLA, Math, and Science operational tests*. https://www.hawaiipublicschools.org/Reports/KAEO%20Technical%20Report_2018_Operational%20Tests.pdf

- Heinrich, C. J., Meyer, R. H., & Whitten, G. (2010). Supplemental education services under No Child Left Behind: Who signs up, and what do they gain? *Educational Evaluation and Policy Analysis*, 32(2), 273–298.
- Hemelt, S. W., & Jacob, B. (2020). How does an accountability program that targets achievement gaps affect student performance? *Education Finance and Policy*, 15(1), 45–74.
- Hemelt, S. W., & Marcotte, D. E. (2013). High school exit exams and dropout in an era of increased accountability. *Journal of Policy Analysis and Management*, 32(3), 323–349.
- Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education*, 20(1), 101–126.
- Hernandez, D. J. (2012). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation*. Annie E. Casey Foundation.
- Hernández, R. A. (2013). *Maintaining a focus on subgroups in an era of Elementary and Secondary Education Act waivers*. Campaign for High School Equity.
- Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academies Press.
- Hill, H. C. (2001). Policy is not enough: Language and the interpretation of state standards. *American Educational Research Journal*, 38(2), 289–318.
- Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. (2006). Using value tables to explicitly value growth. In R. Lissitz (Ed.), *Longitudinal and value-added models of student performance* (pp. 255–290). JAM Press.
- Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*. Harvard University Press.
- Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, 26(4), 11–20.
- Ho, A. D. (2008). The problem with “proficiency”: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, 37(6), 351–360.
- Ho, A. D. (2014). Variety and drift in the functions and purposes of assessment in K–12 education. *Teachers College Record*, 116(11), 1–18.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.
- Hoff, D. J. (2001). New standards’ leaves legacy of unmet goals. *Education Week*, 20(43), 1.
- Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., & Furgol, K. (2011). *Final report on the evaluation of the growth model pilot project*. U.S. Department of Education.
- Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states’ tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* (pp. 163–189). Information Age Publishing.

- Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, 27, 3–17.
- Holme, J. J. (2002). Buying homes, buying schools: School choice and the social construction of school quality. *Harvard Educational Review*, 72(2), 177–206.
- Holme, J. J., Richards, M. P., Jimerson, J. B., & Cohen, R. W. (2010). Assessing the effects of high school exit examinations. *Review of Educational Research*, 80(4), 476–526.
- Hough, H., Byun, E., & Mulfinger, L. (2018). *Using data for improvement: Learning from the CORE Data Collaborative*. Policy Analysis for California Education.
- Howell, W. G. (2015). Results of President Obama's Race to the Top. *Education Next*, 15(4), 58–67.
- Hunter, S. B. (2019). New evidence concerning school accountability and mathematics instructional quality in the No Child Left Behind era. *Educational Assessment, Evaluation, and Accountability*, 31, 409–436.
- Hutt, E., & Polikoff, M. S. (2020). Toward a framework for public accountability in education reform. *Educational Researcher*, 49(7), 503–511.
- Institute of Education Sciences National Center for Education Evaluation and Regional Assistance. (2019). *Presenting school choice information to parents: An evidence-based guide*.
- Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5), 2072–2107.
- Jacob, B. A. (2001). Getting tough? The impact of high school graduation exams. *Educational Evaluation and Policy Analysis*, 23(1), 99–121.
- Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89(5–6), 761–796.
- Jacob, B. (2007). *Test-based accountability and student achievement: An investigation of differential performance on NAEP and state assessments*. National Bureau of Economic Research.
- Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1), 226–244.
- Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, 1(3), 33–58.
- Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118(3), 843–877.
- Jacobs, J. (2015). How is this fair? Art teacher is evaluated by students' math standardized test scores. *Washington Post Answer Sheet*. <https://www.washingtonpost.com/news/answer-sheet/wp/2015/03/25/how-is-this-fair-art-teacher-is-evaluated-by-students-math-standardized-test-scores/>
- Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or shaping public opinion? The influence of school accountability data format on public perceptions of school quality. *American Journal of Education*, 121(1), 1–27.

- Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, 43(8), 381–389.
- Jennings, J., & Sohn, H. (2014). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education*, 87(2), 125–141.
- Jerald, C. D. (2000). The state of the states 2000. *Education Week*. <https://www.edweek.org/education/the-state-of-the-states-2000/2000/01>
- Ji, C. S., Yee, D. S. W., & Rahman, T. (2021). *Mapping state proficiency standards onto the NAEP scales: Results from the 2019 NAEP Reading and Mathematics assessments*. National Center for Education Statistics.
- Jones, L. V., & Olkin, I. (Eds.). (2004). *The Nation's Report Card: Evolution and perspectives*. Phi Delta Kappa International.
- Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies. *Science Education*, 97(4), 621–636.
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Bill & Melinda Gates Foundation.
- Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). National Bureau of Economic Research.
- Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Bill & Melinda Gates Foundation.
- Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, 9(49), 1–22.
- Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47, 180–195.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.
- Koretz, D. (2015). Adapting the practice of measurement to the demands of test-based accountability. *Measurement: Interdisciplinary Research and Perspectives*, 13, 1–25.
- Koretz, D. (2017). *The testing charade*. University of Chicago Press.
- Koretz, D. M., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. RAND.
- Koretz, D., & Beguin, A. (2010). Self-monitoring assessments for educational accountability systems. *Measurement: Interdisciplinary Research and Perspectives*, 8(2–3), 92–109.
- Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). American Council on Education/Praeger.
- Koretz, D., Jennings, J. L., Ng, H. L., Yu, C., Braslow, D., & Langi, M. (2016). Auditing for score inflation using self-monitoring assessments: Findings from three pilot studies. *Educational Assessment*, 21(4), 231–247.

- Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Center for the Study of Evaluation, University of California.
- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5–16.
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253.
- Kraft, M. A., & Gilmour, A. F. (2017). Revisiting “the widget effect”: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234–249.
- Lachlan-Haché, L. (2015). *The art and science of student learning objectives: A research synthesis*. American Institutes for Research.
- Ladd, H. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review*, 18, 1–16.
- Ladson-Billings, G. (1995). But that’s just good teaching! The case for culturally relevant pedagogy. *Theory Into Practice*, 34(3), 159–165.
- Lauen, D. L. (2008). False promises: The school choice provisions in NCLB. In A. Sadovnik, J. O’Day, G. Bohrnstedt, & K. Borman (Eds.), *No Child Left Behind and the reduction of the achievement gap* (pp. 203–226). Routledge.
- Lauen, D. L., & Gaddis, S. M. (2012). Shining a light or fumbling in the dark? The effects of NCLB’s subgroup-specific accountability pressure on student achievement. *Educational Evaluation and Policy Analysis*, 34(2), 185–208.
- Lee, J., & Lee, M. (2020). Is “whole child” education obsolete? Public school principals’ educational goal priorities in the era of accountability. *Educational Administration Quarterly*, 56(5), 856–884.
- Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, 34(1), 109–121.
- LiCalsi, C., Ozek, U., & Figlio, D. (2019). The uneven implementation of universal school policies: Maternal education and Florida’s mandatory grade retention policy. *Education Finance and Policy*, 14(3), 383–413.
- Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives*, 11, 31–31.
- Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind act of 2001. *Educational Researcher*, 31(6), 3–16.
- Liu, J., Lee, M., & Gershenson, S. (2021). The short-and long-run impacts of secondary school absences. *Journal of Public Economics*, 199, 104441.
- Liu, S., Bell, C. A., Jones, N. D., & McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation and Accountability*, 31(1), 61–95.
- Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39(1), 22–52.

- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 44(1), 47–67.
- Loeb, S., & Strunk, K. (2007). Accountability and local control: Response to incentives with and without authority over resource allocation and generation. *Education Finance and Policy*, 2(1), 10–39.
- Lorié, W. (2020). *Addressing the alignment challenge associated with the use of college admissions tests under ESSA*. National Center for the Improvement of Educational Assessment.
- Lyons, S., Johnson, M., & Hinds, B. F. (2021). *Confronting inequity in assessment*. Lyons Assessment Consulting.
- Madaus, G. F., Stufflebeam, D., & Scriven, M. S. (1983). Program evaluation. In G. F. Madaus, D. Stufflebeam, & M. S. Scriven (Eds.), *Evaluation models* (pp. 3–22). Springer.
- Mariano, L. T., & Martorell, P. (2013). The academic effects of summer instruction and retention in New York City. *Educational Evaluation and Policy Analysis*, 35(1), 96–117.
- Mariano, L. T., Martorell, F., & Tsai, T. (2018). *The effects of grade retention on high school outcomes: Evidence from New York City schools*. RAND.
- Marion, S. (2016). *Considerations for state leaders in the design of school accountability systems under the Every Student Succeeds Act*. National Center for the Improvement of Educational Assessment.
- Marsh, J. A., & Farrell, C. C. (2015). How leaders can support teachers with data-driven decision making: A framework for understanding capacity building. *Educational Management Administration and Leadership*, 43(2), 269–289.
- Marsh, J. A., & Koppich, J. E. (2018). *Superintendents speak: Implementing the Local Control Funding Formula*. Policy Analysis for California Education.
- Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education: Evidence from recent RAND research*. RAND.
- Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, 79, 1332–1361.
- McDonnell, L. M. (2004). *Politics, persuasion, and educational testing*. Harvard University Press.
- McDonnell, L. M. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education*, 80(2), 19–38.
- McLaughlin, M. J., & Thurlow, M. (2003). Educational accountability and students with disabilities: Issues and challenges. *Educational Policy*, 17(4), 431–451.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press.

- National Academies of Sciences, Engineering, and Medicine. (2017). *Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress*. National Academies Press.
- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- National Academies of Sciences, Engineering, and Medicine. (2019). *Monitoring educational equity*. National Academies Press.
- National Center for Learning Disabilities. (2018). *Assessing ESSA: Missed opportunities for students with disabilities*.
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. U.S. Government Printing Office.
- National Conference of State Legislatures. (2019). *Third grade reading legislation*. <https://www.ncsl.org/research/education/third-grade-reading-legislation.aspx>.
- National Council on Education Standards and Testing. (1992). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. U.S. Government Printing Office.
- National Council on Measurement in Education. (2019). *National Council on Measurement in Education position statement on the use of college admissions test scores as academic indicators in state accountability systems*. Position statement. https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission_Statement_06-16-19.pdf
- National Governors Association. (1990). *Educating America: State strategies for achieving the National Education Goals*.
- National Research Council. (2011). *Incentives and test-based accountability in public education* (Committee on Incentives and Test-Based Accountability in Public Education, M. Hout & S. W. Elliott, Eds.). National Academies Press, Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education.
- Neal, D. (2013). The consequences of using one assessment system to pursue two objectives. *The Journal of Economic Education*, 44, 339–352.
- Neal, D. (2018). *Information, incentives, and education policy*. Harvard University Press.
- Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92, 263–283.
- NORC. (2015). *Alignment between the 2013 NAEP Grade 8 mathematics assessment and ACT EXPLORE mathematics assessment*.
- Olson, L. (2019). *The new testing landscape: How state assessments are changing under the federal Every Student Succeeds Act*. FutureEd.
- Ou, D. (2009). *To leave or not to leave? A regression discontinuity analysis of the impact of failing the high school exit exam*. Centre for Economic Performance.
- Özek, U. (2015). Hold back to move forward? Early grade retention and student misbehavior. *Education Finance and Policy*, 10(3), 350–377.

- Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1), 163–193.
- Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, 32(1), 5–23.
- Paris, D. (2012). Culturally sustaining pedagogy: A needed change in stance, terminology, and practice. *Educational Researcher*, 41(3), 93–97.
- Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. National Board on Educational Testing and Public Policy.
- Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher*, 39(2), 110–119.
- Petrilli, M. J., Griffith, D., & Wright, B. L. (2016a). *High stakes for high schoolers: State accountability in the age of ESSA*. Thomas B. Fordham Institute.
- Petrilli, M. J., Griffith, D., & Wright, B. L. (2016b). *High stakes for high schoolers: State accountability in the age of ESSA, Part II*. Thomas B. Fordham Institute. <https://eric.ed.gov/?id=ED579520>
- Phillips, M., Reber, S., & Rothstein, J. (2018). *Making California data more useful for educational improvement*. Policy Analysis for California Education.
- Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, 29(4), 3–14.
- Polikoff, M. S. (2012a). Instructional alignment under No Child Left Behind. *American Journal of Education*, 118(3), 341–368.
- Polikoff, M. S. (2012b). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis*, 34(3), 278–294.
- Polikoff, M. S. (2013). Teacher education, experience, and the practice of aligned instruction. *Journal of Teacher Education*, 64(3), 212–225.
- Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal*, 52(6), 1185–1211.
- Polikoff, M. S. (2016, July 12). A letter to the U.S. Department of Education. *On Education Research*. <https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/>
- Polikoff, M. (2022). Alignment. In S. Brookhart (Ed.), *Routledge resources online: Education*. <https://doi.org/10.4324/9781138609877-REE4-1>
- Polikoff, M., & Korn, S. (2020). School accountability. In J. G. Dwyer (Ed.), *The Oxford handbook of children and the law* (pp. 521–550). Oxford University Press.
- Polikoff, M. S., Korn, S., & McFall, R. (2018). *In need of improvement? Assessing the California Dashboard after one year*. Stanford University.
- Polikoff, M. S., McEachin, A., Wrabel, S. L., & Duque, M. (2014). The waive of the future? School accountability in the waiver era. *Educational Researcher*, 43(1), 45–54.

- Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal*, 48, 965–995.
- Polikoff, M. S., & Struthers, K. S. (2013). Changes in the cognitive complexity of English instruction: The moderating effects of school and classroom characteristics. *Teachers College Record*, 115(8), 1–26.
- Polikoff, M. S., & Wrabel, S. L. (2013). When is 100% not 100%? The use of safe harbor to make Adequate Yearly Progress. *Education Finance and Policy*, 8(2), 251–270.
- Polson, C. (2018). TAKS-ing students? Texas exit exam effects on human capital formation. *Economics of Education Review*, 62, 129–150.
- Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, 31(7), 3–14.
- Porter, A. C., Linn, R. L., & Trimble, C. S. (2005). The effects of state decisions about NCLB adequate yearly progress targets. *Educational Measurement: Issues and Practice*, 24(4), 32–39.
- Porter, A. C., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher*, 40, 103–116.
- Porter, A. C., & Polikoff, M. S. (2007, October). NCLB: State interpretations, early effects, and suggestions for reauthorization. *Social Policy Report*, 21(4), 1–15.
- Princiotta, D. (2019). *Understanding the great U.S. high school graduation rate rise: 1998–2010* [Unpublished doctoral dissertation]. Johns Hopkins University
- Quinn, D. M. (2020). Experimental effects of “achievement gap” news reporting on viewers’ racial stereotypes, inequality explanations, and inequality prioritization. *Educational Researcher*, 49(7), 482–492.
- Quinn, D. M., & Desruisseaux, T. M. (2022). Replicating and extending effects of “achievement gap” discourse. *Educational Researcher*, 51(7), 496–499.
- Quinn, D. M., Desruisseaux, T. M., & Nkansah-Amankra, A. (2019). “Achievement gap” language affects teachers’ issue prioritization. *Educational Researcher*, 48(7), 484–487.
- Randall, J. (2021). “Color-neutral” is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, 40(4), 82–90.
- Reardon, S. F., Arshan, N., Atteberry, A., & Kurlaender, M. (2010). Effects of failing a high school exit exam on course taking, achievement, persistence, and graduation. *Educational Evaluation and Policy Analysis*, 32(4), 498–520.
- Reardon, S., Kalogrides, D., Ho, A., Shear, B., Fahle, E., Jang, H., & Chavez, B. (2021). Stanford Education Data Archive.
- Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy*, 6(3), 207–241.
- Rebell, M. A. (2008). *Moving every child ahead: From NCLB hype to meaningful educational opportunity*. Teachers College Press.

- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.) *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37–75). Springer.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417–458.
- Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2), 247–252.
- Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, 2(4), 119–147.
- Roda, A., & Wells, A. S. (2013). School choice policies and racial segregation: Where White parents' good intentions, anxiety, and privilege collide. *American Journal of Education*, 119(2), 261–293.
- Roderick, M., Kelley-Kemple, T., Johnson, D. W., & Beechum, N. O. (2014). *Preventable failure: Improvements in long-term outcomes when high schools focused on the ninth-grade year*. University of Chicago Consortium on School Research.
- Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, 27(4), 309–340.
- Roeber, E., Olson, J., Topol, B., Webb, N., Christopherson, S., Perie, M., Pace, J., Lazarus, S., & Thurlow, M. (2018). *Feasibility of the use of the ACT and SAT in lieu of Florida statewide assessments: Vol. 1. Final report*. Assessment Solutions Group.
- Rose, D. H., Robinson, K. H., Hall, T. E., Coyne, P., Jackson, R. M., Stahl, W. M., & Wilcauskas, S. L. (2018). Accurate and informative for all: Universal Design for Learning (UDL) and the future of assessment. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices* (pp. 167–180). Springer.
- Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). National Center for Research on Evaluation, Standards, and Student Testing.
- Royal, C., & Gibson, S. (2017). They schools: Culturally relevant pedagogy under siege. *Teachers College Record*, 119(1), 1–25.
- Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. *NYU Law Review*, 79, 932–989.
- Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation: Lessons learned from observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research.
- Schneider, M., Teske, P., Marshall, M., & Roch, C. (1998). Shopping for schools: In the land of the blind, the one-eyed parent may be enough. *American Journal of Political Science*, 42(3), 769.
- Shultz, P. K., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education*, 6, 1–11. <https://doi.org/10.3389/educ.2021.701973>

- Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, 152, 154–169.
- Shepard, L. A. (2021). Ambitious teaching and equitable assessment: A vision for prioritizing learning, not testing. *American Educator*, 45(3), 28.
- Shepard, L., Glaser, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement: A report of the National Academy of Education of the NAEP Trial State Assessment: An evaluation of the 1992 achievement levels*. National Academy of Education.
- Shepard, L. A., Marion, S. F., & Saldaña, C. M. (in press). Standards-based reform and school accountability. In L. Cohen-Vogel, P. Youngs, & J. Scott (Eds.), *AERA handbook of education policy research*. AERA.
- Shultz, P. K., & Englert, K. (2021). Cultural validity as foundational to assessment development: An Indigenous example. *Frontiers in Education*, 6, 1–11. <https://doi.org/10.3389/educ.2021.701973>
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.
- Sinnema, C., & Robinson, V. (2007). The leadership of teaching and learning: Implications for teacher evaluation. *Leadership and Policy in Schools*, 6(4), 319–343.
- Skinner, B. F. (1953). *Science and human behavior*. Macmillan.
- Smith, E., & Tyler, R. (1942). *Appraising and recording student progress*. Harper & Row.
- Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Taylor & Francis.
- Smith, S. S., & Mickelson, R. A. (2000). All that glitters is not gold: School reform in Charlotte–Mecklenburg. *Educational Evaluation and Policy Analysis*, 22(2), 101–121.
- Spillane, J. P. (2004). *Standards deviation: How schools misunderstand education policy*. Harvard University Press.
- Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 11(3), 340–359.
- Steinberg, M. P., & Kraft, M. A. (2017). The sensitivity of teacher performance ratings to the design of teacher evaluation systems. *Educational Researcher*, 46(7), 378–396.
- Steinberg, M. P., & Sartain, L. (2021). What explains the race gap in teacher performance ratings? Evidence from Chicago Public Schools. *Educational Evaluation and Policy Analysis*, 43(1), 60–82.
- Stronge, J. H., & Tucker, P. D. (2003). *Teacher evaluation. Assessing and improving performance*. Eye on Education.
- Sun, M., Saultz, A., & Ye, Y. (2017). Federal policy and the teacher labor market: Exploring the effects of NCLB school accountability on teacher turnover. *School Effectiveness and School Improvement*, 28(1), 102–122.
- Toch, T. (1984). School chiefs endorse state-by-state comparisons. *Education Week*, 9, 12.

- Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Education Sector.
- Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education*, 11, 103–126.
- Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.
- Tyner, A., & Larsen, M. (2019). *End-of-course exams and student outcomes*. Thomas B. Fordham Institute.
- U.S. Department of Education. (2005, November 18). *Secretary Spellings announces growth model pilot, addresses chief state school officers' Annual Policy Forum in Richmond* [Press release].
- U.S. Department of Education. (2006, January 25). *Peer review guidance for the NCLB Growth Model Pilot applications*. Washington, D.C.
- U.S. Department of Education. (2009). *Race to the Top Fund. Final rule*. <https://www.govinfo.gov/content/pkg/FR-2009-11-18/pdf/E9-27426.pdf>
- U.S. Department of Education. (2021). *U.S. Department of Education releases guidance to states on assessing student learning during the pandemic*.
- Van Dunk, E., Meissner, D., & Browne, J. (1998). *Parental involvement and school choice: A look at private school choice in Cleveland and Milwaukee*. The Public Policy Forum.
- The Victorian Quality Council. (2008). *A guide to using data for health care quality improvement*. Victorian Government Department of Human Services Rural and Regional Health and Aged Care Services Division.
- Vinovskis, M. A. (2001). *Overseeing the Nation's Report Card: The creation and evolution of the National Assessment Governing Board (NAGB)*. National Assessment Governing Board.
- Walter, M., Kukutai, T., Carroll, S. R., & Rodriguez-Lonebear, D. (2021). *Indigenous data sovereignty and policy*. Taylor & Francis.
- Warren, J. R., & Grodsky, E. (2009). Exit exams harm students who fail them—and don't benefit students who pass them. *Phi Delta Kappan*, 90(9), 645–649.
- Warren, J. R., Grodsky, E., & Lee, J. C. (2008). State high school exit examinations and postsecondary labor market outcomes. *Sociology of Education*, 81(1), 77–107.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Council of Chief State School Officers.
- Webb, N. L. (2002, April 1–5). *An analysis of the alignment between mathematics standards and assessments for three states* [Paper presentation]. American Educational Research Association Annual Meeting, New Orleans, LA, United States of America.
- Webb, N. L. (2005). *Alignment analysis of mathematics standards and assessments: Michigan Grades 3–8*. Michigan Department of Education.
- Webb, N. L. (2010). *Alignment analysis of reading standards and assessments: South Dakota Grades 3–8 and 11, 2009 and 2010*. South Dakota Department of Education.

- Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project.
- West, M. R. (2012). *Is retaining students in the early grades self-defeating?* Center on Children and Families, Brookings Institution.
- Winters, M. A. (2018). *The effect of Florida's test-based promotion policy on student performance prior to the retention decision*. Boston University.
- Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis*, 34(3), 313–327.
- Winters, M. A., & Greene, J. P. (2012). The medium-run effects of Florida's test-based promotion policy. *Education Finance and Policy*, 7(3), 305–330.
- Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, 8(2), 245–279.
- Yeager, D., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). *Practical measurement*. Carnegie Foundation for the Advancement of Teaching.

NOTES

1. Although the statute uses the terminology “valid and reliable academic measure,” the current *Standards* (AERA et al., 2014) and Zwick (this volume) hold that validity is a property of uses of measures and not a property of measures themselves. A valid academic measure is thus one that has sufficient validity evidence to support its intended use in the accountability system.
2. This portion of the chapter is adapted from Polikoff (2022).
3. This portion of the chapter is adapted from Polikoff (2022).

Assessment for Admission, Placement, and Outcomes in Higher Education

Wayne J. Camara

Law School Admission Council

Ou Lydia Liu

ETS

Krista Mattern

NBME

This chapter focuses on three applications of standardized tests in U.S. higher education—use in admissions, placement, and outcomes assessment. Testing often differs across these three applications, but also within each application across institutions based on institutional need and philosophy. In addition, we address the changes to testing and testing policies that have emerged in the past few years related to the pandemic and other externalities.

Admissions tests serve multiple purposes, some of which are explicitly identified by the testing organization, others that may be proposed by the test sponsors or users (e.g., institution, government agencies), and still others that may be stipulated by policy makers and secondary users of test results. Four purposes or claims have been put forth by ACT concerning the use of admissions test scores: (a) admissions (scholarship), (b) placement (course, program), (c) preparation and academic advising (college readiness, academic planning, application to college), and (d) accountability (school, district, and state). College Board lists five similar purposes for SAT scores: (a) students' college/career readiness, (b) admissions and college course placement, (c) student progress, (d) high school placement, and (e) scholarship decisions (Camara et al., 2019). Graduate and professional admissions tests are much more likely to be used only in admissions and scholarship decisions. Admissions test scores have also been used by policy makers and other stakeholders for purposes that are not intended or sanctioned by test developers, such as for preemployment screening, high school graduation, teacher effectiveness, institutional quality, and ranking. In response, the National Council on Measurement in Education (2019) issued relevant cautionary guidance to dissuade such uses. (See also Ho & Polikoff, this volume, for a discussion of the use of admissions tests for nonintended purposes.)

Placement tests have been described as assessments contributing data to institutional and student decisions about course enrollment, course exemption, and credit by examination for students already admitted to a program or institution (Whitney, 1989). Placement tests are used by many colleges and universities, even those institutions with open admissions policies that do not require testing for admissions, to determine which courses a student should initially complete. They include commercial tests and locally developed tests at individual campuses that are most frequently used for placement in math, reading, and writing/composition. Faculty at many institutions have developed placement tests to determine placement in credit versus developmental courses, as well as the appropriate level in sequential coursework such as foreign languages and mathematics.

Assessments of *student learning outcomes* (SLO) are used by higher education institutions for a variety of purposes, including fulfilling regional and program accreditation needs, internal improvement, faculty/staff interest, equity concerns, and responding to accountability calls (Jankowski et al., 2018). The assessment results are typically examined at the group level without significant direct consequences for individual students. A report by the Commission on the Future of Higher Education released under then-Secretary of Education Margaret Spellings increased the demand and prominence of

SLO assessments. The commission's first report (U.S. Department of Education, 2006) noted a lack of mechanisms for the evaluation of outcomes at U.S. institutions, calling for direct and comparable evidence of learning for improved transparency. The most popular assessments include national student surveys, alumni surveys, locally developed surveys, standardized assessments of general knowledge and skills, rubrics, employer surveys, performance assessments, and portfolios.

CHANGES IN ADMISSIONS AND CURRENT CHALLENGES TO ADMISSIONS TESTING

Admissions tests have been shrouded in controversy since the introduction of the SAT (then known as the Scholastic Aptitude Test) in 1926. The debate about innate intelligence versus achievement, large differences in scores between groups (gender, ethnicity/race), and the influence of coaching on test scores are examples of controversies across the decades (Camara, 2009; Donlon, 1984; Jencks & Phillips, 1998; Lemann, 1999; Willingham & Cole, 1997). However, in 2020, a convergence of externalities emerged, which not only disrupted admissions and admissions testing but also appears to have resulted in profound changes to these processes for the foreseeable future.

COVID-19 and the Social Justice Movement

The coronavirus disease (COVID-19), a severe acute respiratory disease, spread rapidly across much of the globe and was responsible for closing U.S. schools as early as March, 2020. It had a profound impact on admissions and admissions testing for the entering class of fall 2021 and beyond. As of April 3, 2020, all schools in 43 states were closed and most schools in other states were similarly closed (*Education Week*, 2020). Given this closure of schools and colleges, administrations of both the SAT and the ACT were canceled, and initial attempts to administer undergraduate admissions tests resulted in significant reductions in capacity because many test centers remained closed and others introduced social distancing protocols well through 2021. The ACT and SAT tested about 50% of students who registered for testing in the summer and fall of 2020 (Jaschik, 2020). Testing companies sought innovative solutions to bring the admissions tests to students in the fall of 2020; the GRE, Graduate Management Admission Test (GMAT), and Law School Admission Test (LSAT) introduced remote proctored examinations by spring/summer 2020, providing access to thousands of students who tested at home. All three graduate admissions testing programs have continued to employ remote proctored exams for testing at home, while the GRE, GMAT, and, most recently, LSAT have also provided candidates a choice of test center administrations (Camara & Mattern, 2022).

The pandemic coincided with the Black Lives Matter movement (Black Lives Matter: About, 2024), and there was a renewed focus and commitment to equity and social justice by colleges, businesses, and many Americans. A survey of 300 top colleges found that the Black Lives Matter movement influenced changes to admissions processes in

about 30% of institutions (Kaplan, 2021). (See also Cook & Pitoniak, this volume, for a discussion of the impact of the Black Lives Matter movement on assessment.)

Before COVID-19 and the increased spotlight on issues of social justice, about 12% of all 4-year colleges or universities neither recommended nor required admissions tests, which represented about 1% of all freshmen in the United States,¹ with one college test blind (i.e., it did not consider test scores for admissions with any applicants; Camara, 2020). For 2024 admission, over 1,900, four-year colleges (85%) did not require the ACT or SAT for admission and 80 colleges were test blind (FairTest, 2024). Graduate admissions and professional admissions tests were also forced to cancel normal administrations of their tests through spring, resulting in a similar movement to test-optional policies among a number of programs. See the following section for a full discussion of test-optional policies. (See also Ercikan & Flores, this volume, for discussion of the impact of the COVID-19 pandemic on assessment.)

Prior to the pandemic, in early 2019, the University of California's (UC) Academic Senate established an 18-person faculty task force to conduct a study of the university's testing practices, focusing on their role in disproportionate disparities in race, ethnicity, and socioeconomic status (SES) evident in the undergraduate population. This was the third time in less than 20 years that the university system had undertaken a review of its admissions policies and practices to ensure fairness and diversity. The task force was charged with addressing the following questions:

- How well do the UC's current standardized testing practices assess entering high school students for UC readiness?
- How well do the UC's current standardized testing practices predict student success in the context of our holistic, comprehensive review process?
- Should UC's standardized testing practices be improved, changed, or eliminated?
- Do standardized tests fairly promote diversity and opportunity for students applying to UC?
- Does UC's use of standardized tests enhance or detract from UC academic eligibility for high school students?

After conducting meetings, inviting relevant experts in admissions and assessment to make presentations, and completing a comprehensive review of research, the task force released its report with six recommendations (Standardized Testing Task Force, 2020). The recommendations generally encouraged UC to consider additional quantitative factors that could supplement high school grade point average (HSGPA) and tests for eligibility, expand the pool of students eligible for admissions based on class rank, expand academic supports for at-risk students, conduct additional research, and explore the development of a new assessment system.

The task force rejected making admissions tests optional, stating, "With some confidence, we can predict that the following would happen if UC stopped using admissions tests and relied solely on (high school) grade point average (GPA) and other aspects of

the student's transcript for the academic appraisal of applicants" (Standardized Testing Task Force, 2020, p. 85):

- The average student admitted to UC would have a lower first-semester grade point average (FGPA), probability of persistence, probability of graduation, and GPA upon graduation.
- The average financial subsidy by the state would have to rise due to the longer time to graduate per student, and costs per bachelor's graduate would rise.
- Comparisons of in-state and out-of-state applicants would be difficult and unconvincing.
- UC would have less ability to target high-risk students for academic support without tests.

The UC task force similarly dismissed a proposal to adopt the Smarter Balanced Assessment Consortium assessment program in lieu of the ACT/SAT, reasoning that test-taker behavior would change if Smarter Balanced were suddenly used for high-stakes admissions decisions and concluding that diversity-related benefits would not be realized. Camara (2023) compared scores for Smarter Balanced and ACT by ethnicity and race and found that, generally, differences were equivalent, meaning use of Smarter Balanced in place of the ACT would not reduce score gaps for underrepresented minorities. The task force (Standardized Testing Task Force, 2020) found little evidence of instructional validity for achievement tests in general and for Smarter Balanced specifically, noting that since the introduction of Smarter Balanced in California, test performance had been flat, suggesting that additional instruction on the content covered by the assessment shows no increases or effect.

Shortly thereafter, UC president Janet Napolitano made several recommendations to the Regents, which they approved:

1. For fall 2021 and 2022 applicants, campuses will have the option to use ACT/SAT scores in selection consideration if students choose to submit them.
2. Beginning with fall 2023 and ending with fall 2024 applicants, campuses will not consider test scores for admissions selection at all and will practice test-blind admissions selection (California students).
3. During the full period of suspension, from 2021 to 2024, students will have the option to submit an ACT/SAT score for use in scholarship consideration and post enrollment course placement, but not admissions.
4. By 1/21, UC will complete a process to identify or create a new test that aligns with the content UC expects students should have mastered to demonstrate college readiness for California freshmen.
5. If UC is unable to either modify or create a test that meets these criteria and can be available for applicants for fall 2025, UC will eliminate the use of the ACT/SAT for freshman admissions (Office of the President of University of California, 2020).

Napolitano's recommendations were clearly at odds with the task force report, but her recommendations also appeared to consider additional threats imposed by other external factors such as COVID-19, the lawsuit brought against UC,² and the persistent and unrelenting challenges to standardized testing at UC. Generally, the UC task force report (Standardized Testing Task Force, 2020) documented evidence that test scores are the single best predictor of academic success at UC; predict well for all major subgroups; are better predictors for underrepresented minorities, first-generation, and low-income students; and had minimal impact on different acceptance rates experienced by subgroups (Camara, 2023).

After the UC Regents adopted a test-optional policy for 2021–2022 and a test-blind policy for 2023, several subsequent events hastened UC's abandonment of test scores altogether. First, a California superior court issued a preliminary injunction stopping the UC system from considering SAT or ACT scores for admissions or financial aid decisions for the class of 2021. The ruling was issued as part of a lawsuit challenging the use of SAT and ACT scores as culturally biased and a violation of equal protection guarantees (*Smith v. Regents of University of California*, 2020). This injunction stated that the test policy adopted by the Regents in the summer of 2020 denied applicants with disabilities equal access to the tests. The ruling reasoned that applicants submitting scores ("submitters") have an inherent benefit denied to nonsubmitters and cited the difficulties of students, particularly those with disabilities requiring accommodations, in registering and completing the ACT and SAT during the pandemic. Next, on May 14, 2021, UC agreed to a settlement with the plaintiffs that made it the first public university system in the nation to stop any consideration of test scores in admissions, scholarship, and financial aid processes (Rippetoe, 2021).

Test-optional policies were clearly gaining in popularity prior to the impact of COVID-19, social justice movements, and UC's task force report. Additional challenges to admissions testing have been attributed to publicized efforts to cheat on tests and gain advantages in admissions, domestically and internationally, and the continued disparity between the demographics of high school graduates and admitted freshmen at many selective institutions. But taken together, these events formed a near-perfect storm, which further fueled alternative admissions processes without test scores.

Test-Optional Policies

Lucido (2018) documented stated and unstated goals of many test-optional policies, which include genuine concerns about the influence and validity of test scores, the desire to enhance the economic and ethnic diversity of the class, and institutional motivations, which may be less altruistic, such as increasing a college's ranking or selectivity (by raising the reported test scores when lower scoring students are nonsubmitters), increasing applications, and increasing yield (since nonsubmitters are more likely to enroll). FairTest (2024) listed over 1,800 colleges and universities that are test optional or test flexible or de-emphasize tests in admissions. Prior to the pandemic, the vast majority of test-optional institutions were private, liberal arts colleges, including 150

for-profit schools and additional institutions with eligibility requirements (e.g., in state, meeting a minimum HSGPA, and submitting other standardized tests). In fact, between 1987 and 2015, only 15% of selective colleges adopted a test-optional policy (Furuta, 2017).

Some institutions that adopted a test optional policy in 2020 have “permanently” moved to test optional with many selective schools continuing their test-optional policies. However, several highly selective institutions announced they were reinstating test requirements for 2025, including Dartmouth, Yale, Harvard, Brown, Caltech, Massachusetts Institute of Technology, Georgetown, Purdue, and the University of Texas at Austin (Hartocollis & Saul, 2024). These institutions cite a variety of reasons for requiring tests, including their own research which shows test scores are the single best predictor of college success at their institutions and the resulting impact when test scores are missing: (a) overreliance on grades with rampant grade inflation (Sanchez, 2023), (b) reliance on the quality and rigor of high schools, (c) rewarding applicants from high schools that regularly send students to the institution and disadvantaging applicants from relatively unknown high schools, and (d) a heightened importance given to personal statements and letters of recommendation – all of which serve to advantage those applicants who are already advantaged (Camara, 2024). Despite uncertainties regarding whether institutions will realize their goals of increased diversity, whether students admitted without test scores perform similarly to students admitted with test scores, and whether undergraduate admissions testing will partially rebound in a post-COVID-19 world, a large majority of institutions are likely to maintain a test-optional policy for the foreseeable future (Camara & Mattern, 2022; Camara, 2024).

Until recently, test-optional colleges had provided only anecdotal evidence that generally claimed there was little to no difference between submitters and nonsubmitters in terms of grades, graduation, or persistence, but evidence allowing comparisons of the two groups in terms of background, high school grades, and other factors was not included (e.g., Hiss & Franks, 2014).

More recently, a few studies of test-optional colleges have investigated these and other issues across multiple institutions. Prior to the pandemic, research showed that among test-optional institutions, 70%–75% of students submitted scores and nonsubmitters were more likely to include a larger proportion of underrepresented racial/ethnic minority students, first-generation students, and low-income students and attain somewhat lower grades in college (Bennett, 2022). One of the first colleges to adopt a test-optional policy was Bates College in Lewiston, Maine, in 1984. Later, a former admission director and colleagues conducted several studies of test-optional colleges. In their most recent study, Syverson et al. (2018) reported that more than half of the test-optional institutions saw larger increases in applications than matched institutions that required testing. However, many test-optional institutions did not see this gain, and the authors failed to explain their criteria and the process used for matching

institutions. The study claimed that test-optional policies did not impact the academic quality of enrolled students and cited a slight increase in the average test score (about 19 points on the SAT scale) and HSGPA (0.02) after going test-optional, but failed to note that test scores were not available for nonsubmitters, who likely had lower scores. The study did report that nonsubmitters had lower high school grades, earned a lower first-semester grade point average (-0.17 on a 4.0 scale), and were slightly less likely to major in science, technology, engineering, and mathematics (STEM), which appears to undermine claims that test-optional policies do not impact academic quality. In a study of 33 schools with test-optional policies, Hiss and Franks (2014) found only marginal differences favoring students who submitted test scores over nonsubmitters in grades (2.88 vs. 2.83 GPA) and graduation rates (64.5% vs. 63.9%), but submitters were more likely to be STEM majors.

Belasco et al. (2015) studied admissions across 180 selective liberal arts colleges, including 32 that were test optional. They found that nonsubmitters were more likely to be from underrepresented and underserved populations. On average, test-optional institutions saw an increase of 26 SAT score points (on a 1,600-point scale), which enhanced their selectivity, and received about 220 more applications than similar institutions that were not test optional; they did not find a difference in ethnic, racial, or economic diversity of enrolled students. Sweitzer et al. (2018) examined the same issues by calculating propensity scores to match test-optional institutions requiring tests, allowing them to model how a test-optional institution would have differed across various outcomes if they had not adopted this policy. Their findings support those of Belasco et al. (2015), with the biggest benefit of test-optional policies being an increase (about 8 points) in SAT scores from a smaller number of applicants and no significant differences in acceptance rates or the diversity of enrolled students. Bennett (2022) compared nearly 100 private institutions that implemented test-optional policies between 2005–2006 and 2015–2016 (a group of earlier adopters) to more than 100 similar institutions that enacted test-optional policies by December 2019 (a comparison group of later adopters). He found enrollment increases for Pell Grant recipients, underrepresented ethnic minorities, and women at test-optional institutions relative to matched peer institutions, but no evidence of increases in applications or yield. Enrollment increases were modest, favoring test-optional schools by 3.1%–4.2% for Pell Grant recipients, but larger for underrepresented minorities (10.3%–11.9%) and women (6.0%–8.0%). Early adopters of test-optional policies did report moderate increases in applications and admission rates, and these benefits may have dissipated as the number of similar institutions with this policy increased over the years. While research seems to clearly suggest that many of the purported benefits associated with test-optional policies are often not supported with data, these policies do not appear to disadvantage institutions.

Data on applications, admissions, and enrollment between 2018–2019 and 2021–2022 from approximately 60 test-optional colleges, representing more-selective public and private 4-year institutions in the United States, were merged with College Board

assessment data to examine how college-going trends and outcomes were affected by pandemic-related disruptions (Edwards et al., 2023; Howell, 2023). The numbers of applications, offers of admission, and enrolled students increased between 2020 and 2021 for nearly all student groups, but more-selective private institutions enjoyed nearly double the increase (29%) in applications compared to all other types of institutions. The proportional representation of ethnic and racial groups changed marginally, while group differences in applications, admission, and enrollment changed very little between 2018 and 2021 among participating institutions. For example, in 2019, the percentage of Black, Native American, and Hispanic/Latino students was 24.4% compared to 25.4% in 2021. The study also found little difference among ethnic and racial groups in their score-sending behavior; students' test scores relative to the college they applied to was the strongest predictor of the decision to disclose their scores (Howell, 2023). In terms of college outcomes, students who disclosed test scores had stronger first-year outcomes (GPA, credits accumulated, retention) compared to students who had no tests or who withheld test scores.

Similarly, Cruce and Sanchez (2023) examined which students were most likely to withhold their ACT scores from test-optional colleges and the impact on persistence to sophomore year. The research focused on matriculants to 142 colleges that introduced a test-optional policy in 2020 and were located in states with an ACT contract to test all public school juniors. The final sample was 102,745 students who attended one of these colleges and graduated from high school between 2017 and 2021. The two predictors examined were (a) discrepancy between students' ACT score and HSGPA and (b) students' ACT score relative to other students' scores at the college. The authors concluded that students make a strategic decision to submit or withhold their test scores, and to help them make that decision, students evaluate their scores against (a) other academic measures (e.g., HSGPA) and (b) the test scores of their peers. Students whose scores are discrepant or who score lower than their peers are less likely to send scores and more at risk of leaving college. Even after all statistical controls were introduced, students who withheld their scores still had a significantly higher likelihood of leaving college within freshmen year or fall to fall.

Chetty et al. (2023) reported that high school grades do a poor job of predicting success at Ivy Plus colleges. Students with a perfect HSGPA of 4.0 achieve a cumulative GPA in college that is less than 0.1 points higher than a student with a 3.2 HSGPA, and higher test scores are associated with higher college GPAs for students from different socioeconomic backgrounds. In addition, students opting to not submit an SAT/ACT score achieve relatively lower college GPAs when they attend an Ivy Plus college.

High School Grades

One important result of test-optional policies is greater reliance on high school grades and other factors. Camara (2018) compared ACT scores of students with an HSGPA of 4.0 or better in the 30 largest schools in a state ($N > 300$ per school) and

reported large between-school differences in the percentage of such students and their mean ACT scores. Schools ranged from having 2.1% to over 23% of students with a 4.0 HSGPA, and average ACT Composite scores differed by over 7 points on a 1–36 point scale. Research on high school grades has revealed systematic grade inflation since the 1990s, which is increasingly beneficial toward wealthy, White, Asian, and private school students (Camara et al., 2003; Hurwitz & Lee, 2018). Consistent with other studies, the UC Academic Senate Task Force (Standardized Testing Task Force, 2020) reported that grade inflation has reduced the predictive power of HSGPA by a third between 2007 and 2015. Chetty and colleagues (2023) reported high school grades have a much weaker association with college GPA. First-year college GPA of students from less advantaged vs. more advantaged high schools, conditioned on SAT/ACT score, were virtually identical, thus displaying no evidence that students from higher-resourced backgrounds outperform students from lower-resourced backgrounds. In addition, reliance on grades alone often results in greater overprediction of college grades for underrepresented minorities than when grades and test scores are used in combination (Mattern et al., 2008; Zwick, 2013), and substantial subgroup differences advantaging White and Asian American applicants on grades remain.

Varsity Blues

In 2019, federal authorities announced charges against over 50 wealthy and famous parents and college personnel in a criminal conspiracy to influence undergraduate admissions at several highly selective colleges, nicknamed Operation Varsity Blues. Parents allegedly paid an admissions consultant between \$200,000 and \$6.5 million to guarantee their children's admission through a variety of means, including bribing college coaches to place students on lists of recruited athletes regardless of their ability or experience, submitting fraudulent documentation of learning disabilities to secure extra time on admissions tests, and bribing test-site supervisors to correct answers on their child's admissions test after completion (Richer & Binkley, 2019). While the scandal impacted all aspects of competitive admissions, it provided an opportunity for some critics to cite another reason why admissions tests may distort the admissions process (Boost, 2019). However, many saw the incident as an extreme case of distorting admissions in illegal ways, an example of what we have long known about the role of affluence in life and education. Joyce Smith, chief executive officer of the National Association for College Admission Counseling, noted that the scandal reinforces much of what we know:

that wealthy people have always enjoyed many legal advantages in the process, such as the ability to pay for tutoring, test preparation, and application coaching, or for the truly wealthy to make sizable donations, fund scholarships or endow buildings or faculty seats. And we know that economic status affects one's educational choices at every stage of life. (Hawkins, 2019, p. 15)

Undergraduate Admissions Practices, Testing, and Legal Challenges

The goals of the admissions process are rarely specified in a manner that facilitates validation. Many examinations of undergraduate admissions have described the goals and desired outcomes as multidimensional as well as disguised (Gaff & Meachum, 2006; Rigol, 2003; Willingham, 1985). Desired outcomes from admissions processes include both individual student outcomes (academic ability, ability to benefit, appreciation of lifelong learning, leadership skills) and institutional outcomes, which may be referred to as “building a class.” Institutions need to ensure diversity and balance along with many factors, such as financial (ability to pay, need- vs. merit-based aid), talents and interests (to support activities, clubs, and majors), and balance (among other factors, including demographics, geography, and socioeconomic factors).

Admissions processes can generally be classified as mechanistic or holistic. Prior to the 2003 Supreme Court decisions in two affirmative action cases involving the University of Michigan, many large, public universities relied on mechanistic or formulaic components in their admissions processes. In many instances, a minimum test score and/or HSGPA, or a combination of both,³ may have been used to reduce the large number of applications to a more manageable ratio (applications:admission slots). Mechanistic approaches and cut scores were efficient tools for institutions with large applicant pools, limited budgets, and the desire to have objective criteria to defend admission decisions to the public. (See Ferrara et al., this volume, for a discussion of cut scores and standard setting.) Conversely, private colleges were more likely to conduct holistic reviews, also known as comprehensive or individual review. The holistic review may differ in how it is implemented, but normally involves the review of all materials, by at least two readers, for each applicant, including the student’s essay and letters of recommendation (Laird, 2005; Rigol, 2004). Additional components of an applicant’s “whole file” may include student accomplishments, opportunities or context, and experiences (e.g., class rank, test score, extracurricular activities, leadership experience, community service). A study examined the relationships between the content of admission essays, self-reported household income, and SAT scores, quantifying essay content using correlated topic modeling and essay style using linguistic inquiry and word count (Alvero et al., 2021). Based on 240,000 admission essays submitted by 60,000 applicants to the UC, they found that essay content and style had stronger correlations to self-reported household income than did SAT scores and that essays explained much of the variance in SAT scores. Overall, there has been little empirical research examining holistic admissions processes, the consistency of ratings attributed to different factors, or the validity of those factors in predicting academic outcomes.

Moreover, an internal audit of undergraduate admissions practice conducted by UC was harshly critical of improper influences and lack of consistent treatment of applicants, finding that campus staff took advantage of weaknesses in admissions processes to admit 64 students as favors to donors, family, and friends. Several weaknesses were

found with the holistic review process, including inadequately trained and supervised reviewers resulting in sometimes overly harsh or overly lenient ratings. They noted that the Office of the President did not review the campuses' admissions processes to detect and prevent unfair or inconsistent practices. Instead, it has allowed weaknesses to persist for years (University of California, Ethics, Compliance, and Audit Services, 2020). Soon after the Supreme Court decision in *Gratz v. Bollinger* (2003), which ruled that the University of Michigan's system of awarding additional points to underrepresented racial/ethnic minority applicants was unconstitutional, many colleges and universities that had not already done so abandoned highly mechanistic admissions processes, and holistic review processes quickly gained popularity. On the same date, the court supported the University of Michigan's Law School in its narrowly tailored use of race in admissions as part of a compelling interest to promote student diversity (*Grutter v. Bollinger*, 2003). To many, a lesson learned from the dual decisions was that race could be considered as part of a holistic review of applicants, but it should not be applied mechanistically with specified points or weights added for minority status. In addition, race should not be a determinative factor in admissions (Rigol, 2004; Zwick, 2018).

Between 1996 and 2012, eight states banned affirmative action in college admissions, and more colleges and universities abandoned strict mechanistic approaches (e.g., cut scores) to make final admission decisions (Davenport et al., 2018). However, the greatest impact from the University of Michigan admissions cases may be the lack of transparency of admissions processes and a decrease in institutional willingness to share data that could disclose subgroup differences among applicants, enrollees, and graduates.

In 2023, the U.S. Supreme Court reversed decades of precedent, effectively ending race-conscious admission programs at colleges and universities. The court found that admissions systems used by Harvard University and the University of North Carolina at Chapel Hill violate the equal protection clause of the 14th Amendment. Chief Justice John Roberts wrote in the majority opinion that "both programs lack sufficiently focused and measurable objectives warranting the use of race, unavoidably employ race in a negative manner, involve racial stereotyping, and lack meaningful endpoints" (*Students for Fair Admissions v. President and Fellows of Harvard College*, 2023, p. 8). The court did say that the decisions should not prohibit colleges from considering an applicant's discussion of how race affected their life in an essay or other means (*Students for Fair Admissions v. President and Fellows of Harvard College*, 2023, and *Students for Fair Admissions v. University of North Carolina*, 2023). In both cases, Asian American students rejected by Harvard University claimed they had been disadvantaged by vaguely defined "personal" ratings that invited implicit bias and discrimination, despite being better qualified on academic measures (test scores, grades, extracurricular activities). The plaintiffs used Harvard's internal admissions data and emails to argue that despite better grades and test scores than applicants from other races, Asian Americans were held to a higher standard and penalized by subjective ratings in the holistic admissions process. Harvard University admitted consideration of race was one factor in their holistic admissions process.

A few days after the court's decision was released, civil rights groups filed a complaint against Harvard University's preferences for legacy applicants with the U.S. Department of Education. The complaint noted that nearly 70% of applicants with family ties to donors or alumni are White and are six times more likely to be admitted than non-legacy applicants. About 28% of Harvard's class of 2019 were legacies, meaning fewer admissions slots were available for non-White applicants, who are far less likely to have family ties to the school (*Lawyers for Civil Rights v. President and Fellows of Harvard College*, 2023).

Prior to these decisions, in 2016, *Fisher v. University of Texas et al.*, the Supreme Court ruled that admissions officials may continue to consider race as one factor among many to ensure a diverse student body.⁴ In writing for the majority, Justice Anthony Kennedy quoted from another landmark decision: "Considerable deference is owed to a university in defining those intangible characteristics, like student body diversity, that are central to its identity and educational mission," adding that the courts must provide universities substantial but not total deference in designing admissions policy (*Fisher v. University of Texas et al.*, 2016, p. 19).

One important element of holistic admissions processes that has generally escaped scrutiny is the reliability and consistency of ratings. If admissions decisions are to be considered predictive of future success, higher education institutions should provide evidence that ratings of each factor (e.g., leadership), element (e.g., admission essay, letter of recommendation), and student record have high interrater agreement. In addition, validity evidence should be presented that demonstrates the admission processes are accurately identifying students who achieve the desired outcomes of the institution. Legal challenges have largely been directed toward the admissions process and decision, rather than the admission tests. However, the impact of these challenges to affirmative action and the increased demand for diversity of admitted students has had a significant impact on admissions tests, which are addressed later in this chapter.

UNDERGRADUATE ADMISSIONS TESTS AND PRACTICES

The SAT

The inception of the College Entrance Examination Board can be traced to a meeting of the Association of Colleges and Preparatory Schools of the Middle States and Maryland held in Trenton, New Jersey, on December 2, 1899. At that time, there was little agreement among colleges about the subject matter content and preparations that should be required of applicants; College Board was founded to bring order to the chaos of different entrance requirements across established Eastern colleges (Donlon, 1984; see also Clauser et al., this volume). The goal for the fledgling organization was to bring standardization and uniformity to entrance requirements for colleges, but it also had the effect of forcing secondary schools to move from a system built on local autonomy to more common curricula. College Board, formally approved on November 17,

1900, was primarily established to serve as a conduit between colleges and preparatory schools to establish uniformity in admissions standards and preparation/curriculum in secondary schools. It achieved these goals in 1901 by establishing requirements for nine subject matter examinations in English, French, German, Latin, Greek, history, mathematics, chemistry, and physics and administering exams to 973 students, which were read by 39 readers with prominence in their field.

Initially, only 3 colleges replaced their own local exams with the new subject tests, but subsequently, an additional 35 institutions agreed to accept the examinations as substitutes for their own requirements. A decade after their inception, in 1911, the number of students separately examined by Harvard, Yale, Princeton, Penn, Bryn Mawr, Stevens Institute, and the Massachusetts Institute of Technology still exceeded the number of students taking the board exams. College Board examinations continued to consist of a series of essay tests requiring readers to grade each exam; but influenced by the efficiencies and success of intelligence tests during World War I and beyond, the College Board added a comprehensive test of general aptitude to its program in 1926—the Scholastic Aptitude Test (Donlon, 1984; Schudson, 1972). The original College Board achievement tests were transitioned to a fully multiple-choice format in 1937 and placed on the same 200–800 SAT scale; they were rebranded as the SAT II Subject Tests in 2005 and then discontinued in 2021.

The 1926 version of the SAT bears little resemblance to the current test, but it marked the transition from essays to multiple-choice-type items on admissions tests. It contained nine subtests, seven with verbal content and two with mathematical content; and time limits were more stringent, with 315 questions administered in 97 minutes. In 1930, the SAT was split into verbal aptitude and mathematical aptitude, a structure that has remained largely in place over time, including the 2024 design (College Board, 2017a; Lawrence et al., 2002). Other major developments that shaped admissions testing occurring in the 1940s included the transition to machine scoring of the SAT, which replaced the monumental task of human scoring of test questions and eliminated the many hours of training clerical workers, as well as the equating of scores on every SAT form to ensure comparability for admissions decisions. ETS (formerly Educational Testing Service) was founded in 1947 by three nonprofit organizations: the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and College Board. ETS was established to assume responsibility for test development, administration, scoring, and research, allowing College Board to continue its role as an association and convener of higher education and secondary schools (Fuess, 1950; Zwick, 2006). The use of the SAT increased dramatically during its first 30 years, with over a half-million students testing annually by the late 1950s and over a million students in 1986.

Changes to the SAT were intermittently introduced in content and question type during its history, but the first major revision of the post–World War II SAT did not come until 1994, when antonyms were dropped from the test, reading passages became longer, calculators were allowed, grid-in items were added in math, and the score scale

was recentered in the summer of 1995 based on a full year of operational scores. During this period, the Scholastic Aptitude Test was renamed the Scholastic Assessment Test and then quickly changed to just the SAT I Reasoning Test. Two additional major revisions of the SAT were introduced in 2005 and 2016, and a digital, multistage, adaptive SAT replaced paper-and-pencil version of the SAT in the United States in 2024.

The major change to the SAT in 2005 largely resulted from 2 years of discussions with the UC and was initiated after a speech by Richard C. Atkinson, then UC president, where he recommended elimination of the test as a UC admissions criterion (Schemo, 2001). The most significant change was the addition of a mandatory writing section composed of multiple-choice items and a single 30-minute essay, resulting in now three separate scores (critical reading, mathematics, and writing), each scored on the 200–800 point scale and resulting in a total score of 600–2400. In addition, analogies were removed as an item type in critical reading, quantitative comparison items were removed in the math section, and additional content was added in advanced math topics to better reflect the curriculum.

The 2016 version of the SAT was introduced in March of that year and represented a major redesign of the previous test in terms of content, design, and the interpretation of scores. The redesign was driven by the adoption of the Common Core State Standards by 43 states and a desire to increase market share through statewide adoption (Dudley, 2016; Felton, 2015). In 2013, just 3 states had adopted the SAT for statewide testing, as opposed to 11 states that were using the ACT. The 2016 version of the SAT returned to issuing two scores—evidence-based reading and writing (ERW; which includes a reading test and a writing and language test) and a math score. The essay, which had been required on the SAT since 2005, became optional. Test takers had 3 hours (plus an additional 50 minutes for the optional essay) to complete the test. The two sections—ERW and math—were reported on the 200–800 SAT scale but required a concordance to compare scores with the prior SAT. Other major changes of the 2016 SAT redesign included the elimination of formula scoring, which imposed a penalty for students' incorrect answers, to rights-only scoring, a simple count of correct items, and the inclusion of items in the math section that prohibit calculator use⁵ (College Board, 2017a). In 2021, College Board discontinued the optional essay, except for school-based administration with state specific contracts (College Board, 2024c).

Three stand-alone field trials were conducted to develop new score scales, equate initial forms to future forms, and build a concordance between the 2005 version of the SAT and the 2016 redesigned SAT; each is described in the technical manual (College Board, 2017a). A single group design was used to establish a concordance by selecting students who completed the 2005 version of the SAT in the fall of 2015 and having them complete the 2016 redesigned SAT in December 2015. The concordance demonstrated that scores on the 2016 redesigned SAT were substantially higher, with the difference favoring the 2016 SAT between 70 and 100 points on the total score scale (400–1600 scale) (College Board, 2016). College and Career Readiness Benchmarks were revised in 2016. Using historical data from the 2005 SAT linked to college outcomes, the SAT

scores associated with a 75% probability of earning a C or higher were derived. Those scores were then transformed to the 2016 SAT score scale based on a 2005–2016 SAT concordance table to set the updated SAT benchmarks at 530 and 480 for math and ERW, respectively (College Board, 2017b).

The SAT was primarily a paper-based test but began offering a digital administration to states and districts in 2017 to assist with comparability studies. The SAT fully transitioned to a digital format in March 2024, which was partially motivated from persistent security challenges encountered by maintaining a large volume paper program internationally. The multistage test design organized content into two stages, each composed of a module of test items comprising half of the section's items. After answering an initial (routing) module of items representing a broad mix of easy, medium-difficulty, and hard items, students are routed to either a lower- or higher-difficulty second-stage module based on their performance on items in the initial module. Additional changes to the test design included a reduced overall testing time from 3 hours to 2 hours, significant shortening of reading passages with one question per passage, and more time per test item. The reading section is composed entirely of discrete items and shorter passages instead of item sets and extended-length passages, while the mathematics section retains multiple-choice items and student produced-response items. PSAT assessments transitioned to digital tests in the fall of 2023 (College Board, 2024c).

In 2022, 18,513 students were recruited to take both a digital SAT and the previous paper-based SAT within a month. Data from the SAT straight-line concordance study of the digital SAT and paper SAT were used to establish a link between the paper-based version of the SAT and the digital SAT's item response theory (IRT) metric for the Reading and Writing section score, Math section score, and performance categories in eight content domains. Scores were linked using the equipercentile method of linking digital IRT maximum likelihood theta values to the paper-based scale scores (College Board, 2024c; see Moses, this volume, for a discussion of scaling, equating, and linking methods).

The ACT

The ACT (then known as the American College Testing) was first introduced in November 1959 in the Midwest region of the United States and was designed with three purposes in mind—admissions, advisement, and placement. The ACT was designed to compete with the SAT, whose primary mission was selection of the few for highly selective institutions (Lemann, 1999). Lindquist (1959) argued that the primary need was “one of finding a type of test that will not just serve a single well-defined purpose, but . . . one of building a multiple purpose rather than a single purpose test” (p. 105). He further argued that such a test should be used more broadly by public institutions and expanded to serve “as the basis for advisement and placement, not just admissions” (ACT, 2009, p. 9). At the time, the SAT remained a tool of primarily selective private institutions. A number of states operated their own statewide assessment system, and there was an absence of a national admissions test, which could be used for students applying to out-of-state institutions. Lindquist argued for the expansion of admissions

testing across the country to a greater number of colleges and well beyond the purview of just elite colleges on the East Coast and to address the needs currently met by a series of tests sponsored by individual states and institutions. Finally, he argued that an admissions test should be based on achievement in curriculum found most often in secondary schools rather than aptitude or innate ability (ACT, 2009; Lindquist, 1959).

The creation of the ACT resulted from both the transformation across higher education in the 1950s, which saw a 433% increase in college enrollment and dramatic increases in educational expenditures, and frustration in attempts to bring significant changes to the SAT (ACT, 2009). Designed by Lindquist and Ralph Tyler to address these purposes, ACT scores were reported to students as well as to colleges; SAT scores continued to be reported only to colleges (Jacobsen, 2017).

The ACT was originally composed of four sections, mirroring the format and content of the Iowa Tests of Educational Development and standard college curriculum: English, mathematics, social studies, and natural sciences. In its first testing year, 132,963 students took the ACT, and that number jumped to nearly 1 million by 1967–1968. At that time, strong regional differences emerged, with 85% of institutions in the Mountain West and Plains states participating in the ACT program compared to 10% for the SAT; conversely, about 17% and 53% of institutions in the northeast and middle states were affiliated with the ACT and SAT, respectively.

Aided by concordance tables first produced in the early 1990s, which allowed admissions officers and high school counselors to compare and convert to a single scale, most colleges accepted scores from either test (Marco et al., 1992). Concordance tables have been produced jointly by ACT and College Board after major redesigns of the SAT in 1994, 2005, and 2016.⁶ Similarity of content and strong statistical relationships are required to support the use of concordances to satisfy different goals, such as determining comparable scores, aligning scales, and predicting a student's unavailable ACT or SAT score from their SAT or ACT score (Dorans et al., 1997). The 2018 ACT–SAT concordance tables represent the last time both tests used matched data to establish score comparisons and includes cautions and guidance to assist users in interpreting scores across tests, although recent changes to both the ACT and SAT seem to require a new concordance. In addition, both testing programs have developed assessments for earlier grades (Grades 8–10), which closely mirror the content and statistical blueprints of the ACT and SAT and are vertically scaled to allow interpretation of growth. In particular, the SAT Suite of Assessments includes the SAT, PSAT/NMSQT, PSAT 10, and PSAT 8/9; ACT assessments include the ACT, PreACT (10th grade), and PreACT 8/9.

In 1989, the natural sciences section was transitioned to the science reasoning section and social studies was changed to a reading section (ACT, 2009). In 2005, ACT was also caught up in the debate initiated by Atkinson and the UC and added an optional 30-minute essay. For the 2005 version, ACT consisted of four sections—English, mathematics, reading, and science, with an optional essay. Section scores, a composite score, and a STEM score are reported on the same 1–36 scale used since its inception. The composite score was the average of English, mathematics, reading, and science;

the STEM score is the average of the mathematics and science scores. In 2005, and later updated in 2013 to reflect a more current reference group, ACT established college-readiness benchmarks by empirically deriving the ACT section score associated with a 50% probability of attaining a B or higher in first-year courses in related subject areas (Allen, 2013; Allen & Sconing, 2005), as well as a more rigorous benchmark for readiness in STEM (Radunzel et al., 2015).

For each ACT section test, the construct that is assessed is considered unidimensional—English, mathematics, reading, and science proficiency. In support of this claim, confirmatory factor analyses based on operational data suggest that a one-factor solution has a good model fit for each of the four subject tests (e.g., Li et al., 2017). In 2021, ACT began to also provide an ACT Superscore, which is calculated by averaging students' best section scores across all test attempts to create a new composite score (Radunzel & Mattern, 2020b). The ACT began offering the test digitally, as well as on paper, for school-based administrations (contracts with states, districts, or schools for census testing during a school-day) in spring 2015 and transitioned all international administrations to digital delivery in fall 2018.

In April 2024, ACT announced it was partnering with Nexus Capital Management LP, a Los Angeles–based private equity firm, and transitioning to a for-profit company. The partnership will fund a nonprofit organization conducting services for education and work success and should provide needed capital to stabilize current offerings and grow new products (Arundel, 2024). In February 2024, ACT began offering a digital administration alongside the paper-administration at selected locations⁸ and introduced other changes for online test takers in April, 2025, which included making the science section optional (as is the writing section), fewer items, more time per item, reduced total testing time, and a reduction in the number of answer choices in the math section from 5 to 4. The ACT Composite score is now comprised of only English, math, and reading (requiring only 125 minutes of testing as compared to 195 minutes based on the prior version), while science scores will be reported separately for those who take it.

These changes, which are collectively labelled ACT Enhancements, were introduced for the online national administrations in April 2025 and will be extended to the paper mode in national administrations and all international administrations in September 2025, and finally school-based administrations (under state and district contracts) in spring 2026 (see Table 17.1). ACT notes that the new format is shorter and allows students more flexibility (ACT, 2024b). It has also been noted that students might have been intimidated by the inclusion of science (Knox, 2024).

A linking study was conducted in June 2024 between the Enhanced ACT design and the tradition design with randomly equivalent groups ($N = 7,600$), confirming the factor structure, showing small differences in the score distributions and lower reliabilities for section scores. The ACT Composite score, now comprised of three section scores (English, reading, math), is not strictly comparable with the ACT Composite score prior to 2025 (English, reading, math, and science) due to both the impact of removing science items from the construct, differences in student performance in science relative

Table 17.1 ACT and SAT Content, Testing Time, and Item Counts

Testing Element	ACT (Enhanced Format)	SAT
Reading section	Reading: 40 minutes, 27 scored items, 9 pre-test items	Reading and Writing: 64 minutes (two 32 minute modules). 50 scored items, 4 pretest items (25 scored items, 2 pretest items per module)
Language section	English: 35 minutes, 40 scored items, 10 pre-test items	
Math section	Math: 50 minutes, 41 scored items, 4 pre-test items	Math: 70 minutes (two 35-minute modules), 40 scored items, 4 pretest items (20 scored items, 2 pretest items per module)
Calculator	Permitted	Permitted
Science section	Optional Science: 40 minutes, 34 scored items, 6 pre-test items	Science: NA
Writing	Optional: 40 minutes essay	NA
Score scale	Composite: 1–36 Section: Reading, English, and Math Optional Sections: Science 1–36; Essay 4 domains and Total 2–12	Total: 400–1600 Section: Reading + Writing 200–800, Math 200–800. Four subscores reported for skills under each section. Reading & Writing = Information and ideas, Craft and structure, Expression of ideas, and Standard English Conventions. Math = Algebra, Advanced math, Problem-Solving and Data Analysis, and Geometry and Trigonometry.
Total testing time	2 hours, 5 minutes	2 hours, 14 minutes
With optional essay and Science	3 hours, 25 minutes	NA
Mode and Digital Format	Paper and Digital available for national and school-based administrations. Digital only for international administrations. Linear format.	Digital delivery. Multistage adaptive (1:2 design).
Pretesting	Embedded within section	Embedded within section

Note. For the ACT, the composite score on the 1–36 scale represents a superscore that is based on English, reading, and math scores. Enhancements to the test are taking place on the following schedule: April 2025 (digital national administrations); September 2025 (paper national administrations); and Spring 2026 (school-based administrations). Information for the ACT is from *ACT Assessment Technical Manual*, 2024a; and *ACT Test Enhancements and Changes*, 2024b; ACT. For the SAT, the digital version was introduced internationally in March 2023 and domestically in March 2024, replacing the current SAT with a revised framework. Information for the SAT is from *Assessment Framework for the Digital SAT Suite* (Version 3.01), 2024a, College Board; and *Digital SAT Launches Across the Country, Completing the Transition to Digital and Providing a Simpler Testing Experience for Students and Educators*, 2024b, College Board.

to other sections, and differences in rounding to derive the Composite score. ACT notes that most students will have a new Composite score that is within the standard error of measurement for the Composite score reported prior to 2025 (Allen & Cruce, 2025, p. 25).

Undergraduate Institutions and College-Going

In 2019–2020, there were 3,982 degree-granting postsecondary institutions with first-year undergraduates. Sixty-seven percent of institutions were 4-year institutions and 33% were 2-year institutions. Among 4-year institutions, 29% were publicly controlled, 58% were privately controlled not-for-profits, and 13% were privately controlled for-profits—a segment that decreased by over 50% between 2013 and 2020. About 25.7 million students are enrolled in U.S. colleges, with roughly 33% enrolled in 2-year institutions and 67% enrolled in 4-year institutions (U.S. Department of Education, 2020). In 2019–2020, there were about 300 public and private not-for-profit 4-year institutions with open admissions in the United States,⁹ which indicates that the vast majority of 4-year colleges have some competitive admissions process that considers prior achievement and experience. However, only 75 of these 1,997 institutions examined admit 25% or fewer applicants. Figure 17.1 illustrates that nearly all (86.2%) four-year colleges admitted at least half of all applicants and over half (57.2%) admitted at least three quarters of all applicants (Center for Education Statistics, 2020).

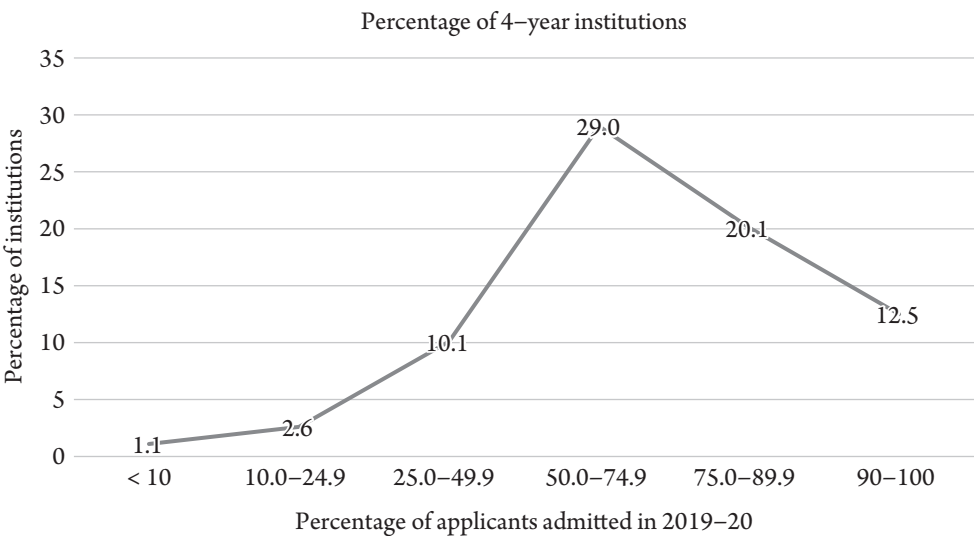


FIGURE 17.1
Percentage of 4-Year Institutions by Undergraduate Admission Rates

Note. The figure does not include the “no application criteria” category, which represents 24.6% of 4-year colleges. Data are from *Digest of Education Statistics, 2020* (NCES 2022-009; Table 219.10), by C. de Brey, A. Zhang, and S. Duffy, 2022, National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Of the 2.9 million high school completers¹⁰ in 2017, about 67% enrolled in college within 12 months of graduation. Since 1985, 4-year college enrollment has remained fairly stable, at 44% of graduates, while 2-year college enrollments increased by a third, to 23% of high school graduates. Immediate college enrollment rates for students from low-income families was 58%, a rate 50% lower than the enrollment of students from high-income families (Camara & Westrick, 2017). Females continue to outnumber males in immediate college enrollment from high school (72% vs. 61%); and while gaps among racial/ethnic groups have been closing in recent years, discrepancies still existed in enrollment in 2017: 87% of Asian students, 69% of White students, 67% of Hispanic students, and 58% of African American students (U.S. Department of Education, 2019).

An annual survey of admissions trends indicated that applications rose through 2015, with over 80% of students applying to three or more institutions and 35% of students applying to seven or more institutions (Clinedinst, 2019). Despite the media's focus on admissions at the most prestigious universities, only 14% of 4-year institutions accepted less than 50% of applicants in one survey (U.S. Department of Education, 2019), with another survey reporting that figure to be 19% of institutions. While such selective institutions accounted for 37% of all applications, they represented 21% of first-year students (Clinedinst, 2019). Table 17.2 illustrates the prepandemic share of U.S. 4-year institutions, applications, and enrolled students by selectivity.

College admission has remained controversial fodder for the media (Bowen & Bok, 1998; Camara, 2009; Shelton, 1997; Soares, 2012; Zwick, 2002, 2017). Though concerns about the employability of college graduates and their student loan debt have long existed (*The Economist*, 2014), the value of a degree, from both a financial and an individual standpoint, has been well established. Higher levels of educational attainment are associated with reduced unemployment and higher wealth accumulation (Boshara et al., 2015). College degree recipients also engage in more prosocial behaviors, such as volunteering, voting, and participating in political activities (Bowen & Bok, 1998; Goldberg & Smith, 2008).

Table 17.2 Percentage of U.S. Applications and Enrollment by Selectivity at 4-Year Institutions for 2017–2018

Institutional Selectivity	Share of Institutions	Mean Number of Applications per Institution	Share of Applications	Share of Full-Time, First-Time Undergraduates
Accept <50%	19.2	12,492	36.5	21.3
50%–70%	33.4	6,108	31.0	32.3
71%–85%	28.8	5,625	24.6	33.4
More than 85%	18.6	2,771	7.9	13.0

Note. N = 1,599. Adapted from *State of College Admission* (p. 8, Table 2), by M. Clinedinst, 2019, National Association for College Admission Counseling.

The numbers are even more dramatic when recipients of advanced degrees are examined. T. Julian (2012) estimated full-time employees with a bachelor's degree would earn approximately \$2.4 million (in 1999 dollars) throughout their adult life compared to \$2.8 million for employees holding a master's degree, while those with a doctorate (\$3.5 million) and professional degrees (\$4.2 million) were estimated to earn significantly more. Given these benefits, it is not surprising that more students choose to attend higher education institutions after high school and that many aspire to earn graduate and professional degrees. Among students completing an admissions test in 2019, approximately 33% aspired to earn a bachelor's degree, and over 43% planned to earn a graduate or professional degree (ACT, 2019a; College Board, 2019a).¹¹

Use of the ACT and SAT in Undergraduate Admissions

A wide variety of measures are considered in undergraduate admissions decisions. Grades and strength of curriculum were of considerable or moderate importance for nearly 90% of institutions, followed by admissions test scores at 83% of institutions prior to the pandemic (Clinedinst, 2019). Additional factors such as writing samples, a student's demonstrated interest, letters of recommendation, and extracurricular activities were less important to most institutions. Scores on state assessments were of no importance to over 70% of institutions. As a result of the pandemic and other external factors impacting admissions, as discussed earlier, national data concerning trends in admissions and the importance of various factors have not been updated recently.

Admission test scores provide normative- and criterion-referenced score interpretations, both of which are used in admissions. At institutions where the number of applicants exceeds openings, applicants' performance on admissions factors, including test scores where required or when submitted, are used to sort between students in a typical norm-referenced fashion. Today, test-optional policies reduce reliance on scores to rank order students or establish a minimum score for entry. Previously, students with higher test scores, higher grades, and more rigorous high school coursework were relatively advantaged in the admissions process over students with lower scores, grades, or course rigor (Standardized Testing Task Force, 2020). The issue is not whether a student will go to college, but whether the student will gain admission to more competitive colleges (Beatty et al., 1999). Colleges with higher ratios of applicants to openings may focus more on selecting students into their institutions, but even less-competitive institutions apply normative admissions processes when they must determine which small percentage of applicants is not admitted. Admissions tests may be used in a more criterion-referenced manner if institutions segment applicants into categories such as admit, deny, and further review or if used along with high school grades in a compensatory manner or a sliding scale such as that employed previously for National Collegiate Athletics Association eligibility in Division I sports.¹²

Test Scores and College Search

ACT and SAT scores may be used to advise students, families, and educators prior to and during the transition from high school to college and by colleges to recruit and build their entering class. Test scores can help students and families reduce the number of colleges and universities under consideration from the nearly 4,000 possibilities to a manageable few by identifying potential institutions where they may have a strong academic match (in addition to satisfying other college preferences). One frequent admissions strategy is to identify potential institutions of interest that fit into three categories in terms of probability of admissions (safe, match, and reach colleges; Hoxby & Avery, 2013).

Accountability

Admission test administrations are increasingly funded by states with a variety of goals, such as identifying students who meet empirical benchmarks for college readiness, increasing participation in postsecondary education, and accountability purposes (Camara et al., 2019). In 2018, 26 states funded the administration of the ACT and/or the SAT for all public school 11th graders (i.e., census testing), and half of those states proposed to use scores for federal accountability under the Every Student Succeeds Act (U.S. Department of Education, 2015), which is not without controversy (Marion & Domaleski, 2019; refer to Ho & Polikoff, this volume, for a more detailed discussion). Many states fund either the ACT or the SAT for all juniors, but a few states allow each district to select among admissions tests. State funding of admissions testing began in 2001 with Illinois and Colorado administering the ACT to all juniors. In 2019, about 43% of admissions testing volumes resulted from state funding of admissions testing (Camara et al., 2019; College Board, 2019a). By 2022, 22 states were using an admissions test for high school accountability (Education Commission of the States, 2022).

States view census testing with the ACT or SAT as an opportunity to get their students on the radar of colleges and to get colleges on the radar of all students, especially among underrepresented groups (Anderson, 2014). Benefits associated with using the ACT and/or SAT as compared to a customized state assessments include increased motivation or effort by test takers (Finn, 2015), providing all students with scores accepted for college admissions, increased college enrollment and access (Hyman, 2017; Klasik, 2013), greater security, and reduced testing time (Camara et al., 2019). Several studies have been published that show statewide testing increases overall or 4-year college enrollment in a state, increases access for underrepresented groups, and is a cost-effective intervention (Allen, 2006; Hyman, 2017; Klasik, 2013).

Participation in Admissions Testing

ACT testing volumes increased significantly since the mid 1990's and surpassed the SAT in total U.S. test takers from 2012 until 2017 (Adams, 2017), when the 2016 redesigned SAT regained several state contracts. In 2019, the last year prior

to COVID-19, approximately 1.8 million and 2.2 million high school seniors had taken the ACT and SAT, respectively. Overall, undergraduate admission testing volumes declined by 14% and 22% for the SAT and ACT, respectively, between 2019 and 2023, hitting a low mark for both programs in 2021 and rebounding slightly thereafter. Camara (2024) documented the increase in applicants withholding test scores from 2019–2020 (22.6%) to 2023–2024 (53.2%). Between 2017–2018, over 50% of high school graduates completed the ACT; and between 2018–2020, more than 50% of graduates completed the SAT. Note that totaling percentages from the ACT and SAT may exceed 100% as some students take both tests. In 2021–2022, participation rates fell precipitously with a larger recovery in SAT test takers by 2023.

It is difficult to estimate the number of U.S. high school graduates who took either admission test, both tests, or neither test because test-taker data, understandably, are strictly held by each testing organization. However, the last official concordance between ACT and SAT in 2018 provides some insight. Table 17.3 uses the projected number of U.S. high school graduates and national data reported by ACT and College Board to estimate the percentage of students completing both, only one, or neither test in 2017. Results from the ACT and College Board concordance were based on 589,753 students graduating in 2017 who were matched across both testing programs (College Board & ACT, 2018). For the current analysis, students who took both tests but did not meet the inclusion criteria for the concordance study were retained, resulting in a slightly larger sample size of 593,500, as reported in Table 17.3. In 2017, about 16.5% of graduates took both tests and about 12.5% of graduates took neither test. More recently, both

Table 17.3 Estimated Number and Percentage of U.S. High School Graduates Taking College Admissions Tests in 2017

Admissions Test	Number of High School Graduates	Percentage
SAT only	1,121,981	31.2
ACT only	1,436,538	39.9
Both tests	593,500	16.5
Neither test	450,928	12.5
Total	3,596,650	100

Note. The total number of U.S. high school graduates projected by the National Center for Educational Statistics for 2016–17, and national data reported by ACT and College Board were used to estimate the percentage of students completing both, only one, or neither test in 2017. The estimate of 593,500 students for the “both tests” value is based on a total of 599,274 students matched across both tests, minus 6,253 international students. The final concordance sample is based on 589,753 students, which excludes students for reasons not relevant to the issue of multiple versus single test participation. In addition, the estimates of students taking both tests are slightly underreported using the concordance samples because they included students who only took the SAT prior to March of their junior year, when a revised SAT was released. Students who took both tests, but one of them was prior to March of their junior year, are counted as taking only one of the tests. Data for projected number of high school graduates are from *Digest of Education Statistics 2017* (NCES 2018-070; p. 193, Table 219.10) by T. D. Snyder, C. de Brey, and S. A. Dillow, 2019, National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Data for students taking both tests is from *Guide to the 2018 ACT/SAT Concordance*, 2018 (p. 1, para. 3), College Board & ACT.

testing programs have witnessed significant decreases in the number of tested students, coinciding with disruptions in testing due to the pandemic along with a sharp uptake in test-optional admission policies. In particular, the 2021 ACT-tested high school graduating class was comprised of 1.3 million students—which is 375,000 fewer students than the 2020 class (Camara & Mattern, 2022). Similarly, there were nearly 700,000 fewer students in the 2021 SAT-tested high school graduating class, which included 1.5 million students as compared to 2.2 million in 2020. Furthermore, the percentage of college applicants submitting test scores has decreased strikingly from 78% in 2019–2020 to 48% in 2021–2022 (Magouirk et al., 2023).

GRADUATE AND PROFESSIONAL ADMISSIONS TESTS AND PRACTICES

Graduate school and professional degree admissions encompass many academic and professional disciplines (e.g., psychology, economics, English literature, medical school, law school) and a range of degrees (e.g., masters, PhD, MD, JD), housed in large public universities, smaller private and public colleges, and public, private, and not-for-profit stand-alone institutions. This chapter will focus on the admissions processes and tests most widely recognized across competitive academic programs and the largest professional schools—the GRE, GMAT, LSAT, and MCAT. Many additional admissions tests are used for entry to specialty programs in medical fields (e.g., dentistry, optometry, pharmacy, veterinary medicine) but fall outside the scope of the current chapter.

The GRE and GMAT have been administered online and in a computer-adaptive fashion since the 1990s (Graduate Management Admission Council, 2020; Wendler et al., 2014). The MCAT and LSAT transitioned to linear computer-based administrations in 2007 and 2016, respectively (Bingham et al., 2012; Margolis, 2018). In addition to providing an opportunity for improved test security and measurement precision, online testing also allows for the inclusion of more innovative item types and more flexibility in test administration models (e.g., remote proctoring).

Institutions responding to an annual survey of graduate school enrollment and degrees reported approximately 2.1 million applications for admission in fall 2020, with about 68% of applications to master's degree and other (e.g., graduate certificate) programs (Zhou & Gao, 2021). Acceptance rates were 22.3% of doctoral applicants and 55.7% of master's/other applicants, with the largest numbers of doctoral applications in social and behavioral sciences, engineering, and biological and agricultural sciences and those of master's applications in business, mathematics, and computer sciences. Approximately 509,999 first-time graduate students enrolled in fall 2020, with 60.8% female students. Unsurprisingly, given the COVID-19 situation, international graduate first-time enrollment was down by 37.4% between fall 2019 and fall 2020, and a large decline was observed in fields such as engineering, mathematics, computer sciences, and physical and earth sciences (Zhou & Gao, 2021).

Undergraduate admissions processes and tests have received their share of scrutiny in both scholarly research publications and popular media. By comparison, graduate school admissions and tests have been subject to much less examination. A report by the Council of Graduate Schools (Kent & McCarthy, 2016) found that graduate admissions processes are primarily decentralized, with over 75% of academic programs responsible for their own admissions program, in contrast to undergraduate admissions, which are generally centralized within an institution and coordinated by admissions professionals. Graduate admissions decisions are often made by small departments or program-level committees and may include an individual faculty member accepting specific students for doctoral-level programs (Michel et al., 2019). Admissions policies and selection rates differ across disciplines and programs.

In recent years, graduate programs have shown a growing interest in applying holistic review to admission decisions, which suggests that the admissions process includes a broad range of factors (Posselt, 2016). Holistic admissions practices hold great potential in helping programs achieve a targeted balance among admitted student cohorts, particularly in terms of student diversity. Relatedly, there has been a good deal of research suggesting that noncognitive measures provide additional predictive value above cognitive assessments while at the same time increasing the diversity of admitted students (see Shultz & Zedeck, 2011, for research in the law school context). Holistic review is considered a useful, effective, and defensible solution for improving diversity following *Grutter v. Bollinger* (2003); however, decentralized admissions processes present unique challenges with implementation, given limited staff and faculty time. The dominant role played by faculty, the large number of applications, the amount of time required for holistic review, and the long-standing culture have all been cited as reasons that holistic review in graduate admissions often relied heavily on test scores (Kent & McCarthy, 2016).

Admission to medical schools and many law schools remains highly competitive, with the most selective schools exceeding a rate of 20 applications per opening. Test scores and undergraduate grades are typically heavily weighted, among all factors considered in admissions. Many programs compute an index of test scores and grades to aid in ranking candidates, but other factors are also considered in final admissions decisions, including grades in college major, personal statement, letters of recommendation, and, in some instances, personal interviews (Zwick, 2006). Note that personal interviews are common in admissions to medical school but are often optional or not employed in admission to other programs.

Admissions Tests for Graduate and Professional Schools

A persistent criticism of graduate admissions is overreliance on test scores and the use of minimum cut scores to advance in the applicant review process. In a review of admissions processes across 10 highly ranked programs, Posselt (2016) disclosed that the influence of test scores extends far beyond what most departments would

admit or what the testing organizations advise. The Council of Graduate Schools found that minimum test scores are widely used “to eliminate a group of applicants before an initial application review . . . a practice largely rumored but sparsely documented” (p. 8).

A few major trends have occurred in graduate school admissions testing since 2010 that are worthy of mention. First, there is an increased call for holistic reviews, which include cognitive and noncognitive factors in admissions, coupled with a desire to increase the diversity of admitted students (Michel et al., 2019). Second, an increasing number of graduate programs have adopted a test-optional policy. Finally, ETS has actively promoted the use of the GRE as an alternative to the GMAT and LSAT in admissions to business programs and law schools, respectively (ETS, 2017, 2020a). Many business schools and some law schools have moved to accept GRE scores in lieu of GMAT and LSAT scores in an effort to attract more nontraditional applicants (Jaschik, 2018; Olson, 2017). Over 55 U.S. law schools accept the GRE in addition to the LSAT for admission, and more than 90% of business schools are estimated to accept either the GRE or the GMAT for admission. Despite the willingness to accept both scores, there is an absence of research examining institutional preferences and use of scores, score comparability, and comparisons of admission rates and success across testing programs. ETS has produced comparison tools to convert GRE scores to both GMAT and LSAT scores; information on sample size and methodology used in each comparison tool was described by ETS (2017, 2020a; as well as by Klieger et al., 2018). However, the Graduate Management Admission Council (GMAC) and the Law School Admission Council (LSAC) have argued that such score conversion tools are flawed (Han, 2022).

Medical and law school applications represent much smaller cohorts, with approximately 55,000 applicants and 23,000 matriculants to U.S. medical schools in 2023–2024 (AAMC, 2023a, 2023b). First-year enrollment across the 203 American Bar Association–approved schools in 2022 was about 38,000, reflecting a decrease from a peak of more than 50,000 students in 2010. There were 62,545 applicants to law school in 2022, which reflects a decrease of about 25,400 applicants from 2010 (LSAC, 2024).

Tests used for admission to graduate and professional schools have a singular purpose—to predict academic success (grades, graduation). Test scores are also used for selection in fellowship programs, for merit scholarships, and by applicants as a guide in selecting among competitive programs. ETS states that scores from the GRE can be used to identify which applicants are academically prepared for graduate-level study and to measure skills required in graduate and professional schools, including business and law. Inappropriate uses of test scores are generally identified by each testing program and include credit by examination, degree conferral, employment decisions, and use as an outcome measure (ETS, 2022). Recently, a number of law schools have withdrawn from the U.S. News rankings. Table 17.4 provides a comparison of the content, sections, scores, score scale, timing, and number of items across graduate and professional admissions tests.

The GRE

The GRE General Test (hereafter referred to as the GRE) is a cognitive test used for admission to graduate schools in the United States and some other countries (ETS, 2023b). It consists of three measures: verbal reasoning, quantitative reasoning, and analytical writing. Verbal reasoning measures the ability to

- analyze and draw conclusions from discourse; reason from incomplete data; identify author's assumptions and/or perspective; understand multiple levels of meaning, such as literal, figurative and author's intent
- select important points; distinguish major from minor or irrelevant points; summarize text; understand the structure of a text
- understand the meaning of individual words, sentences and entire texts; understand relationships among words and among concepts

(ETS, 2023b, Verbal Reasoning section)

with an emphasis on complex verbal reasoning skills. Quantitative reasoning measures the ability to “understand, interpret and analyze quantitative information, solve problems using mathematical models, and apply the basic concepts of arithmetic, algebra, geometry and data analysis,” with an emphasis on quantitative reasoning skills. Analytical writing measures “critical thinking and analytical writing skills, including the ability to articulate and support complex ideas with relevant reasons and examples, and examine claims and accompanying evidence,” with an emphasis on analytical writing skills.

Table 17.4 Graduate and Professional Admissions Tests

Test	Sponsor	Test Sections and Scores (Score Scale)	Description of Test Sections	Notes
Graduate Management Admission Test (GMAT) Focus Edition (Total Score =205–805 ^a)	Graduate Management Admission Council	Data Insights (60–90)	Graphic interpretation, table analysis, multisource reasoning, two-part analysis, integrated reasoning and data sufficiency question types (20 questions, 45 minutes)	The sections are adaptive at the item level. Candidates may select section order. Candidates may bookmark questions in a section for later review; up to 3 answers may be changed. Administered at test centers or at home via remote proctoring.
		Quantitative (60–90)	Problem-solving in applying algebraic and arithmetic knowledge (21 questions, 45 minutes)	
		Verbal (60–90)	This section measures the ability to read and comprehend written material and to reason and evaluate arguments (23 questions, 45 minutes)	

Test	Sponsor	Test Sections and Scores (Score Scale)	Description of Test Sections	Notes
GRE General ^b (Total score = 260–340)	ETS	Analytical Writing (0–6)	A writing assignment to analyze an issue (30 minutes)	Verbal and Quantitative sections are adaptive. Candidates can skip questions or mark them for later review. Administered at test centers or at home via remote proctoring. Candidates can choose which administrations to use in sending scores.
		Verbal Reasoning (130–170)	Ability to analyze and draw conclusions from discourse, reason from incomplete data, understand multiple levels of word meaning, and selecting most distinguishing points, etc. (Section 1: 12 questions, 18 minutes; Section 2: 15 questions, 23 minutes).	
		Quantitative Reasoning (130–170)	Ability to understand, interpret, and analyze quantitative information, problem-solving using math models and basic concepts (Section 1: 12 questions, 21 minutes; Section 2: 15 questions, 26 minutes).	
Law School Admission Test (LSAT) ^c (Total Score = 120–180)	Law School Admission Council	Logical Reasoning (2 Sections)	Ability to examine, analyze, and critically evaluate arguments as they occur in ordinary language (24–26 questions, two 35-minute sections)	Linear digital test. Candidates can skip items and return to them later. The test has been administered at test centers or at home via remote proctoring since August 2023 and will continue through June 2026. A single scale score is reported for the total LSAT.
		Reading Comprehension	Ability to read, with understanding and insight, examples of lengthy and complex materials similar to those commonly encountered in law school (26–28 questions in four sets, 35 minutes)	
		Writing	The test also includes a 50-minute proctored, on-demand writing exam required of all test takers taken remotely and separately from the LSAT. The writing sample is included in the applicant's test record; no score is given.	

(Continued)

Table 17.4 (continued)

Test	Sponsor	Test Sections and Scores (Score Scale)	Description of Test Sections	Notes
Medical School Admission Test (MCAT) (Total score = 472–528)	American Association of Medical Colleges	Biological and Biochemical Foundations of Living Systems (118–132)	Solve problems by combining knowledge of biological and biochemical concepts with scientific inquiry and reasoning skills (59 questions, 95 minutes)	Test is 7 hours and 30 minutes long, with breaks. Linear test administered on computers in test centers. Offered on approximately 30 dates.
		Chemical and Physical Foundations of Biological Systems (118–132)	Integrating chemical and physical foundational concepts with scientific inquiry and reasoning skills (59 questions, 95 minutes)	
		Psychological, Social, and Biological Foundations of Behavior (118–132)	Understanding how psychological, social, and biological factors impact health, research methods, and behavior (59 questions, 95 minutes)	
		Critical Analysis and Reasoning Skills (118–132)	Passages and questions measuring ability to comprehend what is read (53 questions, 90 minutes)	

Note. Information for the GMAT Exam is from *Exam Structure, Content, & Features*, 2023; *Scores*, 2025; and *Understanding Your Score*, 2024b; Graduate Management Admission Council. Information for the GRE General Test is from *The GRE Tests*, 2023c, ETS. Information for the LSAT is from *What to Expect Starting With the August 2024 LSAT*, 2023, Law School Admission Council. Information for the MCAT is from *MCAT Exam Scoring*, 2025a, *What's on the MCAT Exam? Exam Overview*, 2025b; and *What's on the MCAT Exam? PDF Outline*, 2025c, American Association of Medical Colleges.

^a The previous GMAT score scale, prior to the introduction of the *GMAT Focus Edition* in late 2023, was 200–800 (Graduate Management Admission Council, 2024b).

^b The GRE design was shortened in September 2023 from 3 hours and 45 minutes to under 2 hours (ETS, 2023c).

^c The LSAT design changed in August 2024 to eliminate Analogical Reasoning and add a second Logical Reasoning section (Law School Admission Council, 2023b).

In the fall of 2023, the GRE shortened its total testing time to under 2 hours from the 3 hours and 45 minutes it had required. One of the salient features of the GRE has been that for the verbal and quantitative measures, it is adaptive by section with an initial section primarily comprised of questions of medium difficulty; performance on the initial section is used to route test takers into one of three sections of low-, medium-, or high-difficulty questions on the subsequent section. This design allows the GRE to be adaptive at the section level, not at the item level.

The Verbal and Quantitative Reasoning scores are each comprised of two separate sections with a total of 27 items each with total testing time at 41 minutes and 47 minutes, respectively. Previously, the GRE had a total of 80 scored items compared to 54 scored items today. Additional testing time was reduced by removing the unscored

or experimental section used to pretest items and eliminating one of two essay task. Today, an Analytical Writing score is reported from a single 30-minute essay task on analyzing an issue, which requires test takers to evaluate the issue and its complexities and develop an argument with evidence and reasons to support their views (Wendler et al., 2014). GRE test details are also summarized in Table 17.4.

The GRE, then known as the Graduate Record Examinations, were first developed in 1936 as a joint effort in higher education by the graduate school deans of Harvard University, Yale University, Princeton University, and Columbia University and the Carnegie Foundation for the Advancement of Teaching. It was used by the University of Wisconsin in 1938, the first public institution to adopt the GRE. By 1948, over 45,000 students applying to 500 institutions had taken the GRE. The GRE remained a project of the Carnegie Foundation until shortly after the establishment of ETS (then known as Educational Testing Service) in 1947. The GRE was changed from a linear test to an item-level adaptive test in 1993, aiming to improve testing efficiency and the accuracy of estimation of test takers' abilities (Wendler et al., 2014). In 2011, the GRE underwent significant changes (Lewin, 2009), including a new 130–170 scoring scale, the elimination of item types such as antonyms and analogies, the addition of an online calculator for the quantitative section, and the replacement of item-level computer-adaptive testing with section-level adaptive testing (ETS, 2020b). These changes were intended to provide a better test-taking experience and enhance test security. Additionally, the new types of questions in the revised format are intended to test the skills needed in graduate, business, and law school programs (ETS, 2020a). Another feature of the GRE is that it allows test takers to skip questions within a section and change answers within a section. Research suggests that, on average, over 95% of test takers made at least one change to their answers in the verbal reasoning and quantitative reasoning sections; and among those who did change their response(s), 83% were able to improve their verbal reasoning scores and 68% were able to improve their quantitative reasoning scores (Liu et al., 2015).

In addition to its use in general graduate admissions, the GRE was used in admission to more than 1,300 business schools around the world and 91 law schools in the United States as of February 4, 2022. On November 19, 2021, the American Bar Association Council voted to “permit law schools to accept GRE tests scores from applicants in lieu of an LSAT score under Standard 503” (American Bar Association, 2021). Research based on 480 students from 12 institutions shows that both the GRE verbal reasoning and the quantitative reasoning measures were statistically significant predictors of first-year GPA and cumulative GPA for masters of business administration programs (Young et al., 2014). Similarly, research based on 1,587 current and graduated law students from 21 law schools suggests that GRE scores predicted first-year law school GPA when used alone, as well as over and above undergraduate GPA (Klieger et al., 2018). As mentioned earlier, the GRE is administered in Prometric test centers or at home via remote proctoring, which is offered both domestically and internationally with only a few exceptions.

Graduate Management Admission Test

The GMAT is a cognitive test used for admissions into business master and management programs, globally. An updated version of the test labeled the GMAT Focus Edition was released in late 2023¹² (GMAC, 2023). It consists of three 45-minute sections: Quantitative Reasoning, Verbal Reasoning, and Data Insights. The Integrated Reasoning and Analytical Writing sections were removed from the test.

The GMAC, owners of the GMAT, began as an association of 9 business schools in collaboration with ETS in 1953 to develop a standardized test to identify and admit qualified applicants into graduate management programs (GMAC, 2020). The following year, the first GMAT (originally named the Admissions Test for Graduate Business Study) was administered with a testing volume of nearly 3,000 test takers. In 2005, GMAC dissolved its partnership with ETS and selected Pearson and ACT as its new test vendors. In 2018, the number of students taking the GMAT annually peaked at over 200,000, with over 7,000 programs and 2,300 business schools using the GMAT for admission decisions (GMAC, 2018), dropping significantly with the onset of COVID.

The GMAT has gone through various revisions over the years (GMAC, 2020). Notably, the Analytical Writing section was added in 1994. A computer-adaptive version of the GMAT was first introduced to the market in 1997. In 2012, the Integrated Reasoning section was added in response to survey results of nearly 750 management faculty stressing the importance of these skills for management professionals (GMAC, 2011).

In response to COVID-19-related test closures, GMAC began offering an at-home administration option of the GMAT in April 2020. Initially offered as a temporary response, GMAC made this a permanent option and reintroduced test center-based administrations.

GMAC introduced the GMAT Focus Edition in late 2023, composed of three 45-minute multiple-choice sections: Verbal Reasoning, Quantitative Reasoning, and a new section, Data Insights, which replaces the Integrated Reasoning and Analytical Writing Assessment (essay) sections and shortens the test by nearly 1 hour (GMAC, 2023). The Data Insights section presents text, numeric, and graphical information in business scenarios and asks test takers to analyze, interpret, and apply the data. Similar to the GRE, the Quantitative Reasoning and Verbal Reasoning sections of the GMAT exam are computer adaptive; however, for this assessment, the adaptive algorithm works at the item level, selecting items based on the test taker's performance after each item. Three section scores are reported on a 60- to 90-point scale and a GMAT total score is reported on a 205–805 score scale in 10-point increments. Table 17.4 provides additional details on section content and scales. Unlike the previous item-level GMAT CAT, the GMAT Focus Edition allows test takers to mark specific questions within a section and return to them later with the option to change up to three question responses per section (GMAC, 2023).

Law School Admission Test

The LSAT was first administered in 1948 to supplement undergraduate GPA in assisting law schools' admissions professionals (Reese & Cotter, 1994). The

test is owned and operated by the LSAC, a nonprofit corporation in Newtown, Pennsylvania, with more than 200 law schools in the United States and Canada as members. ETS was charged with test development, administration, and scoring through 1979 when LSAC brought these activities in-house. The test's content has been largely unchanged for about 30 years, but two major changes were introduced in July 2019. First, the LSAT transformed to exclusive digital administration utilizing Microsoft Surface Go tablets (Margolis, 2018), with nine national administrations annually. Tablets are stored centrally and distributed to university-based test centers prior to the administration. Second, students taking the LSAT were able to complete the required unscored writing sample in a separate and remote administration. The Digital LSAT consists of four 35-minute sections: two logical reasoning sections, one reading comprehension section, and a variable section used to pretest items that are not operationally scored. Each section is composed solely of multiple-choice items, but students now select the correct option by tapping the screen. A 50-minute required and unscored writing sample is administered digitally and remotely, separate from the test administration.

Law school admissions have fluctuated greatly and the number of LSAC test administrations has also varied over the past 30 years, ranging from 101,000 to over 171,000. Between 1987 and 2019, testing volumes changed an average of nearly 7% annually. Volumes decreased significantly in the late 1990s, increasing in 2002–2011 to a high of over 170,000 and decreasing again to just over 100,000 students in 2014–2017 before rebounding in 2018–2019 (LSAC, 2020). Finally test administrations reached a near-record 169,781 in 2020–21 rebounding from Covid which created barriers to accessing all admissions testing (Kelly & Morgan, 2024). LSAC maintains college transcripts, letters of recommendations, resumes, and personal statements and processes academic credentials for an average of 60,000 law school applicants annually (Anthony et al., 2016). LSAC periodically surveys law school faculty to determine what skills are most essential for success in law school classes; the most recent survey found that higher order critical reading and reasoning skills assessed by the test were ranked highest (LSAC, 2019). Questions used on the LSAT emphasize logic and reasoning skills. Unlike many admissions tests, there is a single scale score on the LSAC (120–180 point scale) computed across 75–76 operational items.

During the pandemic, the LSAT eliminated the variable section and transitioned exclusively to home testing, rebranding the test LSAT-Flex and noting that strict score equivalence with scores from the prior LSAT was not established. In August 2021, the test reintroduced the variable section, retained at-home administration, returned to its original name without the Flex designation, and established score equivalence across versions (Camara & Mattern, 2022). Similar to the GMAC and GRE, in 2023, LSAC introduced dual delivery at home with remote proctoring and test centers during seven 1- or 2-day testing windows. In 2023, the Analogical Reasoning Section was replaced with a second section measuring Logical Reasoning as a result of a settlement to a 2019 lawsuit charging the item type disadvantaged a legally blind test taker (Knox 2023);

and in 2024, a new writing sample was introduced but remained unscored. Table 17.4 provides additional information on the LSAT.

Medical College Admission Test (MCAT)

Introduced in 1928 as the Scholastic Aptitude Test for Medical Students, the admissions test for medical colleges was designed to mitigate attrition rates as high as 50% at some medical schools and included eight separate sections, which were combined into a single score (McGaghie, 2002). Renamed the Medical College Admission Test in 1948, the test had four sections (verbal, quantitative, science, and understanding modern society), each reported on a 200–800 scale. Several other revisions to MCAT occurred subsequently, with the latest major changes introduced in 2015. There was a shift in the focus of constructs from what students know to how well they could apply their knowledge, as well as the introduction of new concepts in biochemistry. In addition, a new behavioral and social science section was added to existing sections, and a new section score scale was introduced (118 to 132) to prevent comparisons to the previous version of the test (Association of American Medical Colleges, 2015). Table 17.4 provides additional information on the four sections, section scale, and total score scale. The MCAT requires more than 6 hours to complete in a single sitting at commercial test centers (Bingham et al., 2012).

A survey of medical school admissions officers found that MCAT total score and undergraduate GPA (total and science/math) were the most important and valued predictors for admission, along with personal interviews at the last stage in the admission process; however, admissions in professional schools now consider many nonacademic factors, suggesting a move to a more holistic approach (Monroe et al., 2013). In 2022, the Association of American Medical Colleges introduced a situational judgment test (named the PREview professional readiness exam) for use by several medical schools in the 2022–2023 admissions cycle. PREview is a standardized exam that presents a series of hypothetical scenarios students may encounter in medical school and asks test takers to evaluate the effectiveness of a series of behavioral responses to each scenario associated with eight competencies (e.g., resilience and adaptability, capacity for improvement, service orientation). It is a remote-proctored, online exam, administered separate from the MCAT (Association of American Medical Colleges, 2022).

RELIABILITY, VALIDITY, AND FAIRNESS OF UNDERGRADUATE AND GRADUATE ADMISSION TESTS

Given the high-stakes nature of admission tests, it is essential that the design and development of these assessments are in adherence with professional standards, producing scores that are highly reliable and have strong validity evidence supporting intended uses (American Educational Research Association [AERA] et al., 2014; see also Huff et al., this volume, for discussion of best practices in test development). Undergraduate and graduate programs use admission test scores to understand applicants' preparedness to succeed in their programs, helping to inform which applicants are admitted.

Given this use, evidence based on relations to other variables is a critical component of the validity argument for admission tests. As such, there is a significant body of literature demonstrating the validity of admissions test scores to predict future educational performance, overall and for student subgroups. This section summarizes the psychometric properties of undergraduate and graduate admissions tests as they pertain to reliability, validity, and fairness.

Reliability

Reliability estimates for undergraduate and graduate admission tests are summarized in Table 17.5. As discussed elsewhere, the ACT, SAT, GRE, and GMAT all underwent significant redesigns between 2023 and 2025, which entailed reducing testing time and the number of items, overall and per sections. In many instances, section score reliabilities were reduced from earlier designs in exchange for other benefits (e.g., less testing time, reduced speededness). Both at the undergraduate and at the graduate level, standardized tests used for admissions are highly reliable, with estimates for the total or composite score typically reaching values of .90 or higher, exceeding the minimum reliability thresholds proposed for high-stakes applications in the past (Nunnally, 1978).

Table 17.5 Reliability of Undergraduate and Graduate Admissions Tests

Test	Test Sections	Reliability
ACT	English	.88
	Math	.88
	Reading	.84
	Science	.85
	Composite	.96
SAT	Evidenced-Based Reading and Writing	.94*
	Math	.90*
	Total	.96*
Graduate Management Admission Test (GMAT)	Quantitative	.90
	Verbal	.89
	Data Insights	.86
	Total	.94
GRE	Analytical Writing	.76
	Verbal Reasoning	.87
	Quantitative Reasoning	.93
Law School Admission Test (LSAT)	Total Test Score	.90

(continued)

Table 17.5 (continued)

Test	Test Sections	Reliability
Medical College Admission Test (MCAT)	Biology and Biochemical Foundations of Living Systems	.82–.86
	Chemical and Physical Foundations of Biological Systems	.82–.86
	Psychological, Social, and Biological Foundations of Behavior	.82–.86
	Critical Analysis and Reading Skills	.82–.86
	Total Score	.95

Note. Data for the ACT Test are from *Initial Evidence Supporting Interpretations of Scores From the Enhanced ACT Test* (ACT Research Report, p. 12), by J. Allen and T. Cruce, 2025, ACT. Data for the SAT are based on the version that existed prior to the introduction of the digital SAT in March 2023 internationally and in March 2024 domestically, and are in the *SAT Suite of Assessments Technical Manual Appendixes* (p. 70, Table A-6.2), by College Board, 2017b. Reliability coefficient estimates are not reported in the technical manual for the new digital SAT, but the reliability of linear digital scores is comparable to that those for its multistage tests per the *Digital SAT Suite of Assessments: Technical Manual* (p. 48), by College Board, 2024c. Data for the GMAT are from *GMAT Focus Psychometric Updates* [Seminar presentation], by K. T. Han, 2025, October 17, GMAT Test Preparation Summit, Reston, VA, United States. Data for the GRE are an average of the reliability estimates for all multi-stage tests administered between September 2023 and May 2024 to reflect the reliability of the shortened GRE, and are in the *GRE Guide to the Use of Scores* (p. 34, Table 4A), by ETS, 2024. Data for the LSAT are from *2021–22 Interpretive Guide for LSAT Score Users*, by Law School Admission Council, 2022. Data for the MCAT are from *The MCAT Exam: Year at a Glance*, by Association of American Medical Colleges, 2015.

Reliabilities exceeding .90 maybe unrealistic and unnecessary for section scores since total scores are primarily used in admissions decisions. For more detailed information, consult the references listed in the notes section of Table 17.5. (See also Lee & Harris, this volume, for a discussion of reliability.)

Validity

For each of these testing programs, a substantial collection of research evidence exists demonstrating the relationship between admission test scores and subsequent higher education success. This section summarizes this body of literature, with common themes emerging. First, test scores are predictive of future performance across multiple indicators of success (e.g., first-year grades, cumulative GPA, retention, graduation), though the majority of research tends to focus on first-year GPA. This is largely due to the expediency with which first-year grades can be collected as compared to more distal outcomes, coupled with more standardization in course taking in the first year as compared to later years. That said, it is still critical and common practice to investigate the validity of test scores for predicting additional metrics of student success. Second, the research points to the incremental validity of test scores for predicting success, above and beyond other admission criteria, most notably prior grades. In particular, studies have examined the degree to which test scores add to the prediction of first-year GPA above HSGPA in the context of undergraduate admissions and above undergraduate GPA (UGPA) in the context of graduate admissions. Results highlight the unique information provided by test scores, underscoring the value of taking a multiple-measures approach in the admission decision

process. Third and finally, for graduate programs, the research evidence suggests that test scores are more predictive of graduate performance as compared to UGPA, whereas HSGPA tends to be slightly more predictive of college success as compared to test scores in the undergraduate context. Refer to Table 17.6 for a summary of validity evidence across the testing program for a focal outcome of interest: first-year GPA; a deeper review of existing validity evidence for each testing program is provided next. (See also Lane & Marion, this volume, for discussion of validity and validation.) Most of the evidence described is based on older test designs prior to changes introduced after 2023. Similar data on the shortened ACT, GRE, and GMAT will be forthcoming.

ACT

Numerous large-scale, national studies have been conducted linking ACT scores to success in college, including performance in the first year, credit-bearing courses, first-year GPA, retention, progress toward a degree, cumulative GPA, and degree completion (e.g., Allen, 2013; Radunzel & Noble, 2012; Sawyer, 2010, 2013; Westrick et al.,

Table 17.6 Validity of Undergraduate and Graduate Admissions Tests for Predicting First-Year Grade Point Average (GPA)

Test	Test Correlation	Prior GPA Correlation	Multiple Correlation
ACT	.51	.61	.64
SAT	.57	.54	.66
Graduate Management Admission Test (GMAT)	.48	.30	.53
GRE	.34–.38	.33	n/r
Law School Admission Test (LSAT)	.60	.42	.66
Medical College Admission Test (MCAT)	.57	.52	.65

Note. For ACT and SAT, first-year GPA is freshman year in undergraduate programs. For other tests, GPA reflects first year of graduate or professional school study. ACT correlations are for paper-based version of the test, prior to the 2024–2025 enhancements, and are corrected for indirect range restriction and criterion unreliability values. Data are from “The Road to Retention Passes Through First-Year Academic Performance: A Meta-Analytic Path Analysis of Academic Performance and Persistence,” by P. A. Westrick, F. L. Schmidt, H. Le, S. B. Robbins, and J. M. Radunzel, 2021, *Educational Assessment*, 26(1), p. 43, Tables 2 and 3. SAT correlations are corrected for restriction of range. Data are from *Digital SAT Suite of Assessments: Technical Manual* (p. 172, Table 8.12), 2024c, College Board. GMAT correlations are for the version prior to the Focused Edition, for which data are not available at this time. Data for the prior version are from *GMAT Exam User Guide for Institutions* (p. 9, para. 5), by Graduate Management Admission Council, 2018. GRE correlations are for the prior design, and are corrected for range restriction and criterion unreliability. Only bivariate correlations between individual GRE sections and first-year graduate GPA are reported. Multiple correlations were not reported (n/r). For predicting the unit-weighted composite of cumulative graduate GPA and faculty ratings, the multiple correlation for three GRE scores was .45; adding undergraduate GPA increased the multiple correlation to .50. Data are from “A Comprehensive Meta-Analysis of the Predictive Validity of the Graduate Record Examinations: Implications for Graduate Student Selection and Performance,” by N. R. Kuncel, S. A. Hezlett, and D. S. Ones, 2001, *Psychological Bulletin*, 127(1), p. 169, Table 2, and p. 175, Table 9. LSAT correlations are corrected for multivariate range restriction. Data are from *Summary of 2017, 2018, and 2019 LSAT Correlation Study Results* (2nd ed., p. 24, Table 3.2b), 2023a, Law School Admission Council. MCAT correlations are corrected for range restriction. Data are from “The Validity of Scores From the New MCAT Exam in Predicting Student Performance: Results From a Multisite Study,” 2020, by K. Busche, M. L. Elks, J. T. Hanson, L. Jackson-Williams, R. S. Manuel, W. L. Parsons, D. Wofsy, and K. Yuan, in *Academic Medicine*, 95(3), p. 394, Figure 3.

2015, 2021). For example, Allen (2013) showed that students who earn higher ACT scores are more likely to earn a B or higher in first-year, credit-bearing courses in related content areas. In addition to specific course grades, students with higher ACT scores are more likely to earn higher GPAs in the first year of college (Sawyer, 2010, 2013; Westrick et al., 2015, 2021). In particular, based on nearly 170,000 students attending one of 50 higher education institutions, Westrick et al. (2015) estimated a mean correlation of .51 for ACT Composite score with first-year GPA, after correcting for multivariate restriction of range. In a follow-up study based on the same data set, Westrick and colleagues (2021) demonstrated the unique contribution of ACT scores above and beyond HSGPA in terms of predicting first-year GPA. In particular, they estimated correlations of .51 for ACT, .61 for HSGPA, and .64 for the combination of the two measures for predicting first-year GPA; all reported correlations were corrected for indirect range restriction and criterion unreliability.

A good deal of evidence also exists demonstrating the relationships between ACT scores and more long-term college outcomes, even after controlling for HSGPA (Mattern & Allen, 2016). For example, Radunzel and Noble (2012) showed that students with higher ACT scores were more likely to achieve a higher cumulative GPA (i.e., 3.00, 3.50, and 3.75) and were more likely to earn a college degree within 6 years. Finally, accompanying ACT's announcement that it is exploring additional test features such as superscoring, section retesting, and online testing on national test dates (ACT, 2019b), a handful of validity studies supporting the new test options have been published (Mattern et al., 2018; Mattern & Steedle, 2020; Radunzel & Mattern, 2020a, 2020b). Validity evidence for the ACT Enhancements were not yet available at the time this volume was published.

SAT

Similar to the ACT, countless studies have been conducted evaluating the predictive validity of SAT scores, alone and in combination with HSGPA, in terms of college success (e.g., Berry & Sackett, 2009; Bridgeman et al., 2000; Kobrin et al., 2008; Mattern & Patterson, 2014; Westrick et al., 2019). These studies demonstrate that SAT scores are predictive of both short- and long-term college success, including course grades, first-year GPA, retention, and degree completion, alone and in combination with HSGPA.

With the redesign of the SAT in 2016, College Board has been conducting numerous studies to document the predictive validity of the revised assessment (Marini et al., 2019; Shaw et al., 2016; Westrick et al., 2019, 2020). Based on over 223,000 students attending one of 171 postsecondary institutions, results from the first national validity study of the redesigned SAT found a correlation of .51 for SAT total and .53 for HSGPA for predicting first-year GPA (Westrick et al., 2019). When the two measures were used in combination, the multiple correlation increased to .61, demonstrating the unique predictive power of each measure; reported correlations were corrected for multivariate range restriction. In preparation for the domestic launch of the digital SAT in

March 2024, College Board conducted a concurrent validity study with 11 undergraduate institutions representing nearly 2,000 first-year, first-time students (College Board, 2024c; Marini et al., 2023). Corrected correlations for restriction of range (and uncorrected correlations) between the Digital SAT with first-year GPA were:

- .57 (.39) SAT Total Score with FGPA
- .53 (.32) SAT Reading and Writing Section Score with FGPA
- .55 (.35) SAT Mathematics Section Score with FGPA
- .66 (.46) SAT and HSGPA with FGPA
- .54 (.27) HSGPA with FGPA
- .12 (.19) SAT incremental validity beyond HSGPA with FGPA.

The study indicated that the relationship between the digital SAT and first-semester GPA was comparable to that of the paper-and-pencil version and that the digital SAT adds incrementally to the prediction of college performance above and beyond HSGPA, providing initial validity evidence for the use of the digital SAT for college admissions. Digital SAT Total scores and HSGPA individually have moderately strong, positive relationships with the number of first-year credits earned, with correlations of .45 and .39, respectively, and a multiple correlation of .50 when employed together (College Board, 2024c).

GMAT

Hundreds of studies have been conducted demonstrating the utility of GMAT scores for predicting success in graduate business programs (Kuncel et al., 2007; Oh et al., 2008; Talento-Miller, 2017). For example, based on meta-analytic results including over 400 independent samples and nearly 65,000 students, Kuncel et al. (2007) found that both GMAT scores and UGPA were predictive of graduate school outcomes including first-year graduate GPA, cumulative graduate GPA, and persistence. For the first-year graduate GPA results, they reported correlations of .47 for GMAT Total and 0.35 for UGPA, after applying corrections for direct range restriction and criterion unreliability. They concluded that the combination of GMAT scores and UGPA would likely result in superior predictive power. Based on the same data set, a follow-up study by Oh and colleagues (2008) suggested that the predictive power of the GMAT is even higher, estimating a correlation of .51 between GMAT Total and first-year graduate GPA when applying different restrictions of range assumptions. GMAC (2018) reported similar findings, with a correlation of .48 for GMAT Total and .30 for UGPA, as well as a multiple correlation of .53 when the two measures are used in combination to predict first-year graduate GPA.

More recently, a national validity study including 28 graduate business programs was conducted, and the results aligned with previous findings (Talento-Miller, 2017). In particular, GMAT Total is a strong predictor of graduate GPA (.38) and shows a stronger relationship with graduate performance as compared to UGPA (.32). Moreover, this study provided validity evidence in support of the Integrated Reasoning and

Analytical Writing sections, reporting median correlations with graduate GPA of .27 and .13, respectively. Finally, the combination of the four GMAT section scores and UGPA resulted in the strongest prediction of graduate GPA, with a multiple correlation of .51. Predictive validity for the newly introduced GMAT Focus Edition was not available at this time.

GRE

Many studies have investigated the GRE scores' prediction of important academic outcomes in graduate programs, such as first-semester GPA, first-year GPA, cumulative GPA, progress in the program, comprehensive examination performance, advisor and instructor ratings, research productivity, degree attainment, and time to degree (e.g., Klieger et al., 2014; Kuncel et al., 2001, 2010; Powers, 2004; Powers & Fowles, 2000). For instance, a meta-analysis consisting of 1,753 independent samples of over 80,000 students showed that both GRE verbal reasoning and quantitative reasoning were strong predictors of first-year and cumulative graduate GPA and that the predictive validity of GRE scores was as strong as that of UGPA (Kuncel et al., 2001). Additionally, the study demonstrated the incremental validity of GRE scores above and beyond UGPA in the prediction of cumulative graduate GPA.

Another predictive validity study of the GRE analyzed data from 10 public universities in Florida based on 21,127 master's students and 4,229 doctoral students from 28 disciplinary domains (Klieger et al., 2014). The results showed that GRE analytical writing was as strong a predictor of graduate GPA as verbal reasoning and quantitative reasoning scores. It is unclear how graduate programs use analytical writing scores in admissions, yet research shows that the analytical writing score is a strong predictor of graduate GPA and is also a good indicator of students, likelihood to be in academic peril, defined as having an average C+ or lower GPA (Bridgeman, 2016; Klieger et al., 2014). Several meta-analyses of GRE's predictive validity have reported small to moderate, positive correlations between scores and graduate first-year and cumulative GPA ($.20 \leq r \leq .36$). A more recent meta-analysis concluded that GRE validities have declined slightly over time, with an average correlation between GRE and student outcomes of .22 in 1957 compared to .15 in 2022. Similarly, the average correlation between GRE and GPA decreased from .25 in 1957 to .16 in 2022 (Feldon et al., 2023).

Researchers have also investigated the GRE's use in graduate admission in international contexts. Results based on a leading research university in Singapore revealed that both GRE verbal reasoning and quantitative reasoning scores predicted first-semester and first-year graduate GPA (Liu, Klieger, et al., 2016). They also predicted program standing (students in good standing are those who never received an academic warning, academic probation, or dismissal). For example, the difference in GRE quantitative reasoning score between students in good standing and those not in good standing was as large as .86 standard deviations for master's programs. Another study analyzed a multinational sample of students at

a university in the Netherlands, and the results indicated that the GRE analytical writing score predicts graduate GPA and master's thesis grades above and beyond UGPA (Schwager et al., 2015; Zimmermann et al., 2017). Furthermore, such prediction is independent of students' SES as indicated by the highest parental degree. Additional empirical validation studies on the shortened GRE should be available in the future.

LSAT

Routinely, the LSAC conducts national validity studies examining the relationship between LSAT scores and first-year performance in law school, alone and in combination with UGPA (Anthony et al., 2013, 2016; LSAC, 2023a; Stilwell et al., 2007, 2009, 2011). Based on the data from the most recent national validity study, which included 163 schools that participated in LSAT Correlation Studies in the 2019 cycle, LSAC reported a correlation of .40 for LSAT and .26 for UGPA in terms of predicting first-year law school GPA (2023a, p. 24, Table 3.2a). When combined, the multiple correlation of LSAT scores and UGPA with law school GPA increased to .49. After correcting for restriction of range, the correlation between LSAT and first-year law school GPA increased to .60; a similar jump was seen for UGPA (.42). A corrected multiple correlation between LSAT and UGPA with first-year law school GPA of .66 was reported (p. 24, Table 3.2b). Consistent with previous LSAC validity studies, the results highlight the superior predictive power of LSAT scores as compared to UGPA and indicate that the combination of the two measures results in the best prediction of law school performance.

MCAT

A good deal of research has been conducted demonstrating the value of MCAT scores for predicting important outcomes for medical students, including grades earned in medical school, performance on the United States Medical Licensing Examination, and unimpeded progress toward graduation (Donnon et al., 2007; Dunleavy et al., 2013; E. R. Julian, 2005). Across these various operational definitions of medical school success, results of these studies have consistently found that MCAT Total scores are predictive of medical school performance and that they have larger validity coefficients as compared to the results based on UGPA. Moreover, MCAT Total scores provide incremental validity above and beyond UGPA, with the combination of the two resulting in the best prediction of medical school performance.

Similar to several of the other testing programs, the MCAT went through a major revision, requiring the collection of new validity evidence to support the use of the revised assessment for medical school admissions. In 2013, a concurrent pilot study was conducted with 11 participating institutions, providing initial evidence of the predictive validity of the new Psychological, Social, and Biological Foundations of Behavior section (Association of American Medical Colleges, 2015). The results

indicated that the new section correlated more highly with medical school grades (.41) as compared to scores from the old exam (ranging from .12 to .30). Busche et al. (2020) conducted an initial validity study of the new 2015 MCAT assessment, examining the validity of MCAT scores, alone and in combination with UGPA, in terms of predicting first-year medical school performance and on-time progression to the second year. The study included roughly 1,000 students who took the new MCAT and entered medical school in 2016 at one of the 16 participating institutions. The results found MCAT Total scores to be predictive of first-year medical school GPA, with a median correlation coefficient of .57 as compared to .52 for UGPA. Based on the combination of both measures, the multiple correlation was .65, highlighting the unique predictive power of both measures. Reported correlations were corrected for range restriction.

Fairness

The fairness of standardized tests is often heatedly debated in the popular press and is the focus of much research of admission tests (refer to Zwick and Rodriguez & Thurlow, this volume, for a more in-depth discussion of fairness in educational measurement and score interpretation). The belief that standardized tests are biased tends to revolve around the finding that there are large performance differences by demographic characteristics (e.g., race/ethnicity, SES). This is true for all testing programs (Association of American Medical Colleges, 2015; ACT, 2024; Busche et al., 2020; College Board, 2017a; Dalessandro et al., 2014; ETS, 2020b; Talento-Miller, 2017), as well as for other measures of educational achievement, such as grades and educational attainment level. That said, subgroup differences, in and of themselves, do not establish test bias (AERA et al., 2014). However, when such findings are observed, follow-up analyses should be conducted to determine what may be contributing to these differences. Subgroup differences by ethnicity/race or SES “do not originate with college admissions tests but are instead reflective of long-standing, systemic variation in educational resources and opportunities that also manifest themselves in achievement disparities on standardized tests at the K–12 level” (Bennett, 2022, p. 183).

More globally, admission testing programs tend to follow a similar protocol to ensure the fairness of their tests (e.g., ACT, 2024; College Board, 2017a; GMAC, 2018; Wendler et al., 2014). This includes providing candidates with appropriate testing accommodations for special populations, conducting extensive item and form reviews by independent fairness experts, analyzing items for differential item functioning, and examining test data for subgroup differences, differential validity, and differential prediction. More detailed research on the fairness of admission tests with a focus on differential validity and prediction results is described below for each testing program. Most data are based on older test designs prior to changes introduced after 2023.

ACT

Differential validity and prediction by student subgroups have been the focus of much research on the ACT (e.g., Allen, 2016; Radunzel & Noble, 2013; Sanchez, 2013). For example, Sanchez (2013) examined the degree to which the ACT exhibited differential prediction of first-year GPA by race/ethnicity, gender, and income level. Results showed that the strength of the relationship was similar across subgroups, with slight overprediction for underrepresented racial/ethnic minority students, males, and low-income students. Based on the same student subgroups, Radunzel and Noble (2013) reported similar results when examining the differential prediction of more distal college outcomes, including cumulative GPA and degree completion. A subsequent study by Mattern et al. (2017) found that the underprediction of female performance was reduced roughly by half after controlling for conscientiousness. Contrary to the belief that test scores are biased against underserved students given lower average scores, research consistently shows that underserved students tend to perform worse in college than predicted based on their ACT scores.

SAT

Differential validity and prediction results for the SAT mirror the ACT findings. In particular, the strength of SAT scores for predicting important college outcomes, including first-year GPA, cumulative GPA, and graduation, is similar across student subgroups (Marini et al., 2019; Mattern et al., 2008; Mattern & Patterson, 2014; Young, 2004). SAT scores tend to slightly overpredict college performance for underrepresented racial/ethnic minority students, students from lower SES backgrounds, and males; the extent of overprediction associated with HSGPA exceeds that of SAT scores for underrepresented racial/ethnic minority students. Various explanations for differential prediction have been offered, such as differences in course-taking behavior and noncognitive factors (Keiser et al., 2016).

GMAT

The degree to which the relationship between GMAT scores and graduate school performance varies for relevant student subgroups has been the focus of numerous research studies. For example, in their 2007 meta-analysis, Kuncel et al. tested whether the relationship between GMAT scores and first-year graduate GPA was moderated by gender, native language, full-time status, and undergraduate major. The results indicated that the predictive strength of GMAT scores did not vary by student subgroups except for a slightly stronger relationship for part-time students.

More recently, Talento-Miller (2017) examined the differential validity and differential prediction of GMAT scores in terms of graduate GPA by gender, race/ethnicity, citizenship, native language, and age. Based on data from 28 business programs, the results indicated that the predictive strength of GMAT Total scores is similar across subgroups, with slightly higher correlations for older students (≥ 31) and slightly lower

correlations for non-native English speakers and students who were citizens of East or Southeast Asia. In general, the results support the predictive validity of GMAT scores, overall and for student subgroups.

GRE

Prior research has also examined GRE's prediction by common subgroups such as gender and ethnicity. For example, Burton and Wang (2014) examined GRE's prediction of first-year GPA and faculty ratings in the mastery of the discipline, professional productivity, and communication skills across five disciplinary areas (biology, chemistry, education, psychology, and English). Results did not show consistent over- or underprediction for gender groups. Also, after course taking was controlled within a discipline, the gender prediction differences largely disappeared. The findings showed sizable overprediction for African Americans. However, the authors speculated that the sample size for this group may have had an impact on such conclusions. Kuncel and Hezlett (2007) reported that underprediction for females was observed at the undergraduate level but not at the graduate level and cited differential disciplinary choice and course selection as possible reasons accounting for the differential gender prediction.

LSAT

In addition to examining the overall validity of LSAT scores for predicting first-year GPA among law students, the LSAC also regularly investigates the degree to which LSAT scores exhibit differential prediction of first-year law school GPA by relevant subgroups, alone and in combination with UGPA. Two reports were recently published investigating this issue by gender and racial/ethnic subgroups using data from schools participating in the LSAT Correlation Studies from 2012 to 2015 (Rahman & Trierweiler, 2017; Trierweiler & Rahman, 2017). Encompassing data from 148 law schools with sufficient racial/ethnic sample sizes, the results indicated the LSAT first-year GPA validity coefficients were similar across racial/ethnic subgroups (Trierweiler & Rahman, 2017). Both LSAT scores and UGPA overpredicted first-year GPA for underrepresented racial/ethnic minority students, with the combination of the two measures resulting in the least amount of differential prediction.

As for the analyses by gender (Rahman & Trierweiler, 2017), the results were based on data from 158 law schools. Similar to the racial/ethnic findings, the results indicated the LSAT first-year GPA validity coefficients were similar for males and females. In terms of differential prediction, LSAT scores, alone and in combination with UGPA, exhibited minimal prediction error by gender, supporting the validity and fairness of LSAT scores for use in law school admission decisions.

MCAT

Prior studies have examined the degree to which MCAT scores exhibit differential validity, with results indicating comparable validity across racial/ethnic subgroups (Busche

et al., 2020). However, with the introduction of the new MCAT in 2015, the collection of evidence supporting the validity and fairness of new MCAT scores is essential. In response to that research gap, Busche and colleagues (2020) not only investigated the overall predictive validity of the new MCAT scores but also examined whether new MCAT scores exhibited differential prediction of first-year medical GPA and on-time progression to Year 2 by demographic characteristics (gender, race/ethnicity, and highest parental education level). Based on an analysis of mean residuals, there was no evidence of differential prediction by the subgroups examined; none of the mean residuals were significantly different from zero.

PLACEMENT TESTS

Placement Decisions: Remediation Versus Credit-Bearing Courses

According to the *Standards* (AERA et al., 2014), placement tests are designed to be efficient and usually cover only a subset of prerequisite content taught in high school. Postsecondary institutions may use scores to determine student eligibility to enter credit-bearing courses rather than remedial developmental courses or to determine the appropriate initial course in a sequenced area of study (e.g., English, foreign language, mathematics).

Historically, placement in remedial or college-level courses has been based on a single placement test score, often locally determined through a study or faculty judgment. Remedial courses may be helpful to some students, but they also require students to make a substantial investment of time and money that could otherwise be applied to college-level coursework. Studies have suggested that the effects of remedial courses on student outcomes are at best mixed for those who are thought to be on the cusp of needing additional academic support. Students who enroll in one or more remedial courses must still pay tuition without earning credits toward graduation and remedial coursework may exhaust students' financial aid before they graduate, reducing their graduation rates (Attewell et al., 2006). Colleges also incur a large financial burden from developmental programs, estimated at more than \$4 billion in 2015 (Rodriguez et al., 2015).

The high cost in dollars and opportunity associated with remedial courses spurred legislation and policies that sought to limit remediation by eliminating remedial courses at some institutions (e.g., 4-year colleges), require evidence of readiness for college admissions (e.g., ACT/SAT or placement test scores), or require the use of multiple measures for placement decisions. While the number and percentage of students placed into remediation show a moderate decrease since 2010, nearly 70% of new students entering a 2-year college each fall are assigned to such programs (Chen & Simone, 2016; Rodriguez et al., 2015).

Bailey et al. (2010) examined the enrollment patterns of 257,672 first-time credential-seeking students who began their enrollment in fall 2003 to fall 2004 in 57 colleges that participated in a developmental program entitled "Achieving the

Dream.” Student enrollments and course completions were tracked for 3 years. The researchers found that about 30% of students referred to developmental education did not enroll in any remedial course, and only about 60% of referred students enrolled in the remedial course to which they were referred. The results also show that more students exit their developmental sequences because they did not enroll in the first or a subsequent course than because they failed or withdrew from a course in which they were enrolled. Bailey et al. (2010) also found that students who enrolled in a college-level class despite a recommended placement in remedial courses were slightly less likely to pass the course than students who were placed directly into the college-level course by virtue of passing the placement test but were much more successful in the course than students who complied with their remedial placement, because relatively few students who entered remediation ever attempted the college-level course. Despite the prevalence of remedial programs in colleges, studies of program effectiveness and the validity of placement tests assigning students to such programs have produced at best mixed results, with most studies suffering from small samples and a lack of rigor (Scott-Clayton, 2012). A more comprehensive study of remediation explored the interaction of remediation and academic preparation, finding that students who exhibited poor precollege academic preparation (as measured by high school grades, the highest math course taken in high school, and college admission test scores) and completed remedial programs fared better on a range of college outcomes than similarly prepared students who did not complete remediation programs; however, these findings did not hold for students who had moderate or strong precollege academic preparation and were assigned to remedial courses (Chen & Simone, 2016).

In 2012, the National Assessment Governing Board commissioned a college survey to determine which placement tests and cut scores were most prevalent. The study documents differences across 2-year and 4-year institutions (Fields & Parsad, 2012). Table 17.7 illustrates that 80% of 2-year institutions use a placement test in math compared to 65% of 4-year institutions. ACT math was used at about a quarter of institutions (2-year and 4-year), whereas the SAT was being used at 20% of 4-year institutions but only 12% of 2-year institutions. Seventy-three percent of 2-year institutions reported using a placement test in reading as compared to less than 40% of 4-year institutions. ACT’s Compass and Asset and College Board’s ACCUPLACER placement tests were the most popular instruments at 2-year colleges, but both ACT tests were discontinued because ACT cited research showing placement tests, in general, are not as effective at determining readiness as previously thought (Fein, 2015). Of course, these data are now over 10 years old, and patterns in usage may have changed.

Given that test scores tend to be only moderately predictive of course success, concerns about the degree to which test scores lead to incorrect placement decisions have surfaced, especially in light of the poor educational outcomes for students placed in remedial college courses (Barnett & Reddy, 2017). Placement tests have been found

Table 17.7 Estimated Percentage of U.S. Postsecondary Institutions Using Selected Math and Reading Tests for Placement in Fall 2011

Institutional Level	Est. Number of Institutions	% Using Any Math Test	ACT Math	SAT Math	ACCUPLACER Elementary Algebra	ACCUPLACER College Math	Compass Algebra	Compass College Algebra	Other Math
All institutions	3,650	71	23	17	16	5	20	4	22
2-year	1,470	80	23	12	24	7	34	8	16
4-year	2,180	65	24	20	11	3	10	2	26
Institutional level	Estimated Number of Institutions	% Using Any Reading Test	ACT Reading	SAT Critical Reading	ACCUPLACER Reading Comprehension	ASSET Reading Skills	Compass Reading	Other Reading	
All institutions	3,650	53	16	11	19	9	22	10	
2-year	1,470	73	21	10	29	19	43	12	
4-year	2,180	39	13	12	12	2	9	8	

Note: The number of institutions may not sum to totals because of rounding or suppressed data. Adapted from *Tests and Cut Scores Used for Student Placement in Postsecondary Education: Fall 2011* (p. 11, Table 1, and p. 17, Table 6), by R. Fields and B. Parsad, 2012, National Assessment Governing Board, U.S. Department of Education.

to underplace students, in comparison to coursework completed and grades, having an adverse impact on underrepresented minorities. In response, an emerging trend in college placement is for institutions to take a more holistic approach (Rutschow & Mayer, 2018), using additional measures such as high school grades, high school courses completed, and performance in dual enrollment or Advanced Placement grades. Moreover, legislative and policy eliminating the use of placement tests have been associated with large increases in retention, transfer credits, and graduation rates at 2-year colleges (Zamundio-Suarez, 2022). About 13% and 21% of all institutions report using one of these measures along with test scores for placement decisions in reading and math, respectively (Fields & Parsad, 2012).

ACCUPLACER

College Board's ACCUPLACER placement testing program is the only remaining major large-scale placement test after ACT discontinued both Compass and Asset placement tests in 2016. For over 30 years, ACCUPLACER (originally branded as Computer Placement Tests) has been used to assess student readiness for college-level courses. The ACCUPLACER battery of placement tests, labeled Next Generation ACCUPLACER, was redesigned in 2016 based on data from a survey of higher education users. New items were written based on new specifications and field tested and calibrated to a new reporting scale, 200–300, with increments of 1. The battery includes tests of English Literacy (EL) and tests of computer skills, but this chapter will focus only on the Next-Generation placement tests: Reading (20 questions), Writing (25 questions), Arithmetic (20 questions), Quantitative Reasoning, Algebra, and Statistics (20 questions), and Advanced Algebra and Functions (20 questions).

ACCUPLACER has also been used with high school students to identify students prepared to enter college credit-bearing courses or in need of further developmental instruction prior to graduation. The technical manual states that scores can be used to identify high school students “eligible” to enroll in dual enrollment courses but provides no evidence or guidance on such uses (College Board, 2019b). All ACCUPLACER assessments are computer adaptive, fixed length, and tailored to each test taker using an item-selection algorithm based on a weighted deviations model.¹³ A number of constraints are built into the assessments, including content constraints. In 2016, 11 million ACCUPLACER tests were administered, but that number counts each assessment completed; the same student will often be routed through multiple tests in a single subject (e.g., all three math or EL tests) to receive a single placement score (College Board, 2017c, 2019b); additionally, volume counts are based on the larger set of assessments used, including Next-Generation, traditional, EL, computer skills, write placer, and diagnostic tests.

The program handbook states, “Multiple factors should be used when making high-stakes decisions like college placement” (College Board, 2018, p. 4) and encourages colleges to conduct a local validity study if establishing cut scores. The Next-Generation

ACCUPLACER redesign incorporated changes in content, item types, and other elements, but publicly available statistical specifications are based on the prior design (College Board, 2017c).

Technical Characteristics

Reliability estimates range from .80 to .92, and standard errors of measurement range from about 7 to 12 points on the 200–300 point scale (L. Bay, personal communication, April 27, 2020). The ACCUPLACER Battery relies on content validity to support claims and provides a detailed description of content and skills measured. Although criterion-related (e.g., predictive) studies have not been reported for ACCUPLACER Next-Generation, the previous battery had moderately strong evidence supported by a meta-analysis of 47 placement validity studies across 17 institutions between 2001 and 2006 (Mattern & Packman, 2009). The mean operational validity coefficient for individual ACCUPLACER tests in the previous battery (before 2016) ranged from .16 to .36 ($M = .24$) when success was defined as obtaining a “B or higher” and from .10 to .32 ($M = .19$) when success was defined as obtaining a “C or higher,” suggesting a weak to moderate relationship between ACCUPLACER scores and course success. However, mean validity coefficients increased to .42 and .34 after correcting for restriction of range. Validity of placement tests is also evidenced by examining the percentage of students recommended for placement in a course who eventually pass that course. Such studies have shown moderate to strong relationships for the previous tests.

STUDENT LEARNING OUTCOMES

Definition of SLO Assessment

SLO have received heightened attention from many higher education institutions for reasons such as accreditation, accountability, and internal improvement (Liu, 2023). The National Institute for Learning Outcomes Assessment (NILOA; 2019), a pioneer in the promotion of valid use and effective practices of SLO assessment, defines learning outcomes as “the expected knowledge, skills, attitudes, competencies, and habits of mind that students are expected to acquire from an educational experience (para. 1).” It is important to note that learning outcomes encompass both domain-specific knowledge and skills and generic, transferable competencies.

In the acquisition of domain-specific knowledge and skills (e.g., journalism or economics), students typically follow a structured progression of a curriculum, designed purposefully to build and reinforce their knowledge and skills in these domains. In this regard, students’ increased understanding and proficiency can be largely attributed to their experience in that defined sequence of courses.

Students’ development of generic, transferable competencies such as critical thinking and intercultural competency is more complicated and can hardly be attributed to a single college experience. For example, institutions may not offer critical thinking

courses but still include critical thinking as a measured outcome, since students could gain critical thinking skills from exposure to a broad range of courses, programs, and institutional activities. Students may hone critical thinking skills, for example, from a philosophy course that discusses inductive and deductive reasoning and from a statistics course that introduces correlation and causation. Students could also gain critical thinking skills from participating in an undergraduate research program or being part of a debate club. Furthermore, they could take a Massive Open Online Course (MOOC) on critical thinking. All these activities can contribute to a student's critical thinking skills without the student's enrollment in a course dedicated to critical thinking. In this sense, students' critical thinking skills reflect the overall learning experience they have at an institution.

In addition to the important distinction between domain-specific and generic learning outcomes, it is also critical to note that the focus of college learning outcomes by higher education institutions is on a spectrum of cognitive abilities, personality and behavioral traits, and social and emotional skills (Pellegrino & Hilton, 2012). The focus is closely aligned with the "whole student" notion and also reflects the demands for both technical and nontechnical skills in the modern workplace. A notable example is the three clusters of competencies identified: cognitive (e.g., critical thinking, creativity), intrapersonal (e.g., intellectual openness, work ethics), and interpersonal (e.g., collaboration, leadership). Moreover, the value placed on nontechnical skills by employers has increased over time. The 2018 IBM Institute for Business Value Global Country Survey revealed that adaptability, time management and prioritization, teamwork, and communication overtook technical skills in becoming the most valued skills by employers (LaPrade et al., 2019).

Notable International SLO Efforts

In parallel to the attention SLO assessment has received in the United States, there are also a number of notable international initiatives about the promotion of college-level learning outcomes. For example, the Assessment of Higher Education Learning Outcomes (Tremblay et al., 2012) project, sponsored by the Organisation for Economic Co-operation and Development (2013), attracted 17 countries to participate in its pilot examining college students' performance in critical thinking, quantitative reasoning, and economics, reflecting a focus on both generic and domain-specific skills.

Another notable example is the Modeling and Measuring Competencies in Higher Education—Validation and Methodological Innovations (KoKoHs) project, funded by the German Federal Ministry of Education and Research over two phases, totaling 9 years (2011–2015; 2015–2019). KoKoHs was focused on the assessment of competencies in both general and disciplinary-specific domains and built a consortium of institutions and programs in Germany and with its international partners in addressing students' learning outcomes (Zlatkin-Troitschanskais et al., 2018).

The European Union also funded a two-phase project titled Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe to identify

a common assessment framework for five broad areas of disciplinary domains: engineering/civil engineering, social sciences/education, humanities/history, health care/nursing, and natural sciences/physics. On the basis of collaborations between faculty members and with input from ETS, a series of assessment frameworks were published articulating the knowledge, skills, and competencies students are expected to master and demonstrate at both the bachelor and the master levels (Wagenaar, 2018).

Different Types of SLO Assessments

The majority of U.S. higher education institutions specify their own learning objectives and student learning outcomes. According to a survey of 811 regionally accredited, undergraduate degree-granting institutions, Jankowski et al. (2018) reported that 82% of institutions have specified learning outcomes. The institutions surveyed reported using multiple assessment approaches to measure learning outcomes using four different types of tools on average. The most popular assessment tools include national student surveys, alumni surveys, locally developed surveys, standardized assessments of general knowledge and skills, rubrics that define competencies (e.g., VALUE Rubrics by Association of American Colleges & Universities), employer surveys, external performance assessments, and portfolios. Notable examples of standardized SLO measures include the HEIghten suite of assessments (Liu et al., 2014), the Collegiate Learning Assessment (Kaniuka & Wynne, 2019), and the California Critical Thinking Assessment (Davies, 2011). The HEIghten assessment suite includes five independent modules in critical thinking, written communication, quantitative literacy, civic competency, and intercultural competency (Griffith et al., 2016; Liu et al., 2014; Roohr et al., 2014; Sparks et al., 2014; Torney-Purta et al., 2015). Rubrics are also a popular tool used by faculty members to evaluate learning outcomes. Each assessment tool has its own advantages and limitations. For example, national student surveys are easy to implement and can be an efficient way of gathering information on student experience and engagement on campus, yet they typically do not provide direct evidence of student learning. Rubrics offer great flexibility for adaptation but could fall short of consistency in implementation if users are not adequately trained. As Liu (2017) pointed out, there is no one-size-fits-all solution when it comes to assessing student learning outcomes. Depending on their assessment goals, priorities, and resources, institutions should select the tools that best fit their needs.

How SLO Assessments Have Been Used and Current State

SLO assessments serve many purposes, such as a program or regional accreditation, internal institutional improvement, faculty/staff interest, equity concerns, and responding to accountability calls (Jankowski et al., 2018). The assessment results are typically reported at an aggregate level. As noted earlier, the purpose of the Commission on the Future of Higher Education was to identify prominent issues in American higher education and develop improvement solutions. The commission's first report (U.S. Department of Education, 2006) sparked a series of accountability initiatives that required participating institutions to provide more data on key performance metrics

to the public, with particular attention to the assessment of learning outcomes using standardized measures. The Department of Education also provided a grant with a component to examine the comparability of students' performance in critical thinking and analytic writing assessed across three standardized measures (Klein et al., 2009). National accountability efforts include the Voluntary System of Accountability for public 4-year institutions and the Voluntary Framework of Accountability created for community colleges.

As is true for many initiatives and changes in education, the focus on accountability has evolved since 2012 as a result of challenges in implementation and shifting policy guidance. The impact of accountability initiatives on institutions' use of learning outcomes assessment has declined significantly, from 31% in 2007 to 13% in 2017 (Jankowski et al., 2018). Internal improvement has become an increasingly important driver for institutions to examine students' learning. Some of the most significant challenges that institutions face include decreasing student enrollment, rising student debt, competition from nontraditional education providers (e.g., online programs, boot camps), and lack of connection to the workforce. When asked whether most Americans have an accurate view of higher education, only 15% of college and university presidents agreed or strongly agreed (Inside Higher Ed, 2019). For many institutions, the focus on SLOs has been driven by the need to prove value instead of a response to federal calls for accountability.

Prominent Issues in Assessing Learning Outcomes

Test-Taking Motivation

In this section, we discuss a number of issues in the implementation and interpretation of SLO assessments. As previously mentioned, most institutions administer SLO assessment to drive institutional change, rather than to monitor individual learning. The assessment outcomes rarely bear direct consequences for students. As a result, students' motivation for taking SLO assessments has been a long-standing concern. Abundant research suggests that when test stakes are low, test takers may not try their best; and therefore, their performance tends to be an inaccurate reflection of their abilities (e.g., Braun et al., 2011; Gneezy et al., 2019; Liu et al., 2012; Wise & DeMars, 2005). Gneezy et al.'s (2019) study revealed that if U.S. students had used their best effort in the Programme for International Student Assessment (PISA), U.S. ranking on the PISA mathematics test would have been improved from 36th to 19th. Researchers have explored strategies to help improve students' motivation. Offering incentives contingent on performance seems to be an effective way of improving motivation and performance (e.g., Braun et al., 2011; Gneezy et al., 2019).

Measuring Performance Differences

A concept that was frequently explored in the beginning of the accountability initiatives around 2007 was value-added learning of college education, implying that the amount of learning that can be attributed to the effectiveness of higher education could be

accurately measured. Although the concept was appealing in holding education providers accountable for the instruction and experiences they provide, there were many designs as well as methodological and practical constraints that prevented meaningful interpretations. In the analysis of change of learning, longitudinal samples are typically preferred over cross-sectional samples because of the invariance in key background variables for longitudinal samples. However, longitudinal samples are very difficult to obtain in higher education because the first-year to second-year attrition rate is about 17% for 4-year public institutions and 38% for 2-year colleges (National Student Clearinghouse Research Center, 2019). In addition, significant transfer rates also make it difficult to track students. About 37% of students who start at 2-year institutions and about 39% of students who start at 4-year institutions transfer to another institution within 6 years (Shapiro et al., 2018). These are the main reasons that many studies employ cross-sectional samples when examining learning gains in college (Klein et al., 2009; Liu et al., 2012; Roohr et al., 2019). However, students in different cohorts may vary in key background and academic variables, which could potentially influence the comparability of these groups (Liu, Liu, et al., 2016). For example, senior students who manage to remain in college may have possessed higher academic qualifications on average when they were freshmen than the current freshmen in the cross-sectional sample, thus introducing inequality in the two samples to be compared. There is another issue of differential participation from freshmen and seniors in that seniors are less likely to sign up for voluntary institutional testing because of other priorities, which further exacerbates the sample inequality. The confluence of methodological challenges, differential motivation (e.g., freshmen try harder than seniors), and mobility in higher education poses difficulty on any attempt to disentangle the contribution of any individual institution on student learning. Researchers have also provided solutions to address unequal samples when examining cross-sectional learning gain. For example, Liu, Liu, et al. (2016) explored the application of a propensity scoring weighting approach aiming to control for possible differences in key background variables (e.g., gender, ethnicity, college entrance exam scores) between cohorts of freshmen and seniors used in the analysis.

Reimagining Learning Outcomes Assessment: Connecting Assessment, Learning, and Careers

Despite the widespread use of learning outcomes assessments, institutions have varying degrees of success in utilizing the assessment results for driving improvement and innovation. A unique challenge around the interpretation and use of generic SLO assessments is that the results are typically not directly related to any course but instead reflect a culmination of a wide range of experiences in and outside college. Institutions at times struggle with identifying a practical entry point to provide remedies when students show a deficiency in a generic field. According to Jankowski et al. (2018), there are many examples of expansive assessment efforts that lead to limited changes on campus. In the following section, we discuss how the use of SLO assessments can be reimagined to help enhance students' key competencies and increase connection with the workforce.

Employers have long been concerned about skills gaps, and such concerns are exacerbated by industry changes introduced by automation. McKinsey Global Institute (2017) projected that by 2030, about 3%–14% of the global workforce will need to switch occupational categories, and all workers will need to adapt as their jobs evolve with automation. An employer survey from the National Association of Colleges and Employers (2018) suggested significant gaps in key competency areas, including critical thinking/problem-solving, global/multicultural fluency, career management, digital technology, leadership, teamwork and collaboration, oral and written communication, and professionalism and work ethic. These skills are also emphasized in Liu, Kell, Liu, et al. (2023) in which the researchers proposed a new vision for assessment that reflects the continuity of skills development throughout K–12, higher education, and workforce. Although institutional assessment reflects a wide range of educational objectives, having a connection to employer expectations has great potential for increasing students' workforce readiness and engaging students in the assessment process because the results will likely bear more significance for students. Offering a clear definition of what these skills mean and how they manifest in concrete contexts is also important as institutions strive to help students improve on these critical skills, which calls for close collaboration between education providers and employers to contextualize these skills (Liu et al., 2021; Liu, Kell, Williams, et al., 2023). Currently, learning resources for skill improvement are provided from a range of sources, including locally or institutionally developed resources MOOCs, and offerings from education start-ups and non-traditional education programs. Making students aware of these resources and offering more systematic support following the assessment activities has great potential to motivate students to capitalize on the learning opportunities for skill improvement.

CONCLUSION

The use of tests in higher education, especially for admissions, remains one of the most highly contested uses of standardized testing in America. College degrees remain a highly valued commodity in our society. Entrance to selective institutions is valued more than entrance to typical or less selective institutions, and the ratio of applicants to openings is highly restricted in the most selective undergraduate and many graduate and professional degree programs. As long as these conditions exist, there will be heated debate about who gets in and the tools and criteria used to make such decisions. As Zwick (2017) reminded us, there is no universally accepted definition of merit or equity and no completely objective way to evaluate admissions systems or decisions. The outcomes of admissions processes serve as *prima facie* evidence of a lack of fairness for many stakeholders, and accumulated evidence of validity, reliability, and fairness of scores is often unpersuasive.

However, results of the body of research on admissions tests continue to reflect their value and efficacy, as described by a 1999 report from the National Academy of Sciences. This study (Beatty et al., 1999) noted that standardized tests provide an efficient, low-cost, valid, and reliable method to identify talent (college readiness) and minimize

differences that exist between schools, grades, and learning opportunities. A more recent study found that admissions tests result in greater diversity, and policies that would eliminate or reduce their use would reduce diversity and have other negative effects on educational outcomes (University of California, Academic Senate, 2020). Although this finding was not directly rejected, the UC president and Regents did reject the recommendations from the task force. However, admission tests must also be used appropriately, which precludes reliance on test scores as the primary or sole measure used for admissions, placement, or assessment of learning outcomes. The discrepancy between the public's perception of standardized tests and the research that contradicts common beliefs points to the need for the measurement community to do a more effective job in promoting a national dialogue of assessment that is data driven and evidence based.

Another important trend emerging in the use of assessments in higher education includes the increasing influence of technology because all major admissions test are now digital and models of testing such as remote proctoring and increased use of artificial intelligence have emerged. In turn, the marked demands for increased flexibility and customization may challenge the very principles associated with standardized testing and score equivalence (Camara & Davis, 2022). Similar to K–12 state assessments in accountability, four of the major admissions tests discussed have also undertaken substantial design changes to reduce testing time and length. Reducing testing time, stress of anxiety of test takers, and providing free and substantial programs for test preparation have gained importance in this industry segment. We should anticipate continued efforts to improve the test taker experience, given its heightened importance in a test optional environment where learners have much greater autonomy and flexibility in their testing choices.

In the brief period between completion of the first draft of this chapter (late 2019) and the final version (2025), testing in higher education has undergone the most profound changes in its history. These changes, continued challenges, and future uses are still very fluid and unknown as we conclude. In the next few years, we anticipate the release of additional research, which can better help us evaluate the efficacy and utility of admissions testing, and we hope such research will also address the efficacy and consequences of admissions practices and decisions made based on other factors. Placement testing has undergone additional scrutiny and the assessment of learning outcomes is in its infancy. The next decade will present challenges to testing organizations and higher education in both continuing practices of assessing students for admissions and placement and deriving meaning from assessments of student learning.

ACKNOWLEDGMENTS

We would like to thank the editors, Linda Cook and Mary Pitoniak, for their guidance and coordination of this work and especially our reviewers, Jerry Lucido, Emily Shaw, and Rebecca Zwick, for their detailed review, suggestions, and comments over multiple reviews, which only strengthened the manuscript.

REFERENCES

- ACT. (2009). *ACT: The first fifty years, 1959–2009*.
- ACT. (2018). *ACT/SAT concordance*. <https://www.act.org/content/dam/act/unsecured/documents/pdfs/ACT-SAT-Concordance.pdf>
- ACT. (2019a). *The ACT Profile Report—National graduating class 2018*. https://www.act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
- ACT. (2019b, October 8). *ACT test to provide new options designed to offer students more choices, faster results*. <http://leadershipblog.act.org/2019/10/act-test-to-provide-new-options.html>
- ACT. (2024a, January). *ACT Assessment technical manual*. https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf
- ACT. (2024b). *ACT Test enhancements and changes*. <https://www.act.org/content/act/en/products-and-services/the-act/test-changes/enhancements.html>
- Adams, C. J. (2017, May 30). In race for test-takers, ACT outscores SAT—for now. *Education Week*. <https://www.edweek.org/ew/articles/2017/05/24/in-race-for-test-takers-act-outscores-sat-for.html>
- Allen, J. (2006). *The benefits of statewide use of the ACT Test*. ACT. <https://files.eric.ed.gov/fulltext/ED510717.pdf>
- Allen, J. (2013). *Updating the ACT college readiness benchmarks* (Report No. 2013-6). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2013-6.pdf
- Allen, J. (2016). *Subgroup representation and validity in ACT College Readiness Benchmark development samples*. ACT. <https://www.act.org/content/dam/act/unsecured/documents/5483-tech-brief-subgroup-representation.pdf>
- Allen, J., & Cruce, T. (2025, June). *Initial evidence supporting interpretations of scores from the enhanced ACT test* (ACT Research Report). ACT. <https://www.act.org/content/dam/act/unsecured/documents/R2425-Initial-Evidence-Supporting-Interpretations-of-Scores-from-the-Enhanced-ACT-Test-2025-06.pdf>
- Allen, J., & Sconing, J. (2005). *Using ACT assessment scores to set benchmarks for college readiness*. (Report No. RR2003-1). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2005-3.pdf
- Alvero, A. J., Glebel, S., Gebre-Medhin, B., Antonio, A. L., Stevens, M. L., & Dominique, B. W. (2021). Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science Advances*, 7(42), 1–10. <https://doi.org/10.1126/sciadv.abi9031>
- American Bar Association. (2021, November 19). *Summary of actions of the section's council at its public meeting Nov. 19, 2021*. https://www.americanbar.org/content/dam/aba/administrative/legal_education_and_admissions_to_the_bar/council_reports_and_resolutions/nov21/21-nov-legal-ed-summary-of-council-meeting.pdf
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Reliability/precision and errors of measurement. In *Standards for educational and psychological testing* (pp. 33–47). <https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf>

- Anderson, N. (2014, August 20). Growing number of states fund ACT college admissions testing for 11th-grade students. *Washington Post*. https://www.washingtonpost.com/local/education/growing-number-of-states-fund-act-college-admission-testing-for-11th-grade-students/2014/08/19/eb0edcba-27a7-11e4-86ca-6f03cbd15c1a_story.html
- Anthony, L. C., Dalessandro, S. P., & Reese, L. M. (2013). *Predictive validity of the LSAT: A national summary of the 2011 and 2012 LSAT correlation studies* (LSAT Technical Report No. TR-13-03). Law School Admission Council.
- Anthony, L. C., Dalessandro, S. P., & Trierweiler, T. J. (2016). *Predictive validity of the LSAT: A national summary of the 2013 and 2014 LSAT Correlation Studies* (LSAT Technical Report No. TR-16-01). Law School Admission Council.
- Arundel, K. (2024, April 18). ACT partners with private equity firm, will transition to for-profit. Higher Ed Dive. <https://www.k12dive.com/news/act-assessment-private-equity-college-admissions/7134450/>
- Association of American Medical Colleges. (2015). *The MCAT exam: Year at a glance 2015*.
- Association of American Medical Colleges. (2022). *AAMC PREview professional readiness exam*. <https://students-residents.aamc.org/aamc-preview/aamc-preview-professional-readiness-exam>
- Association of American Medical Colleges. (2023a). *Table A-10: Applicants to U.S. MD-Granting Medical Schools by Race/Ethnicity (Alone) and State of Legal Residence, 2023–2024*. <https://www.aamc.org/media/6036/download?attachment>
- Association of American Medical Colleges. (2023b). *Table A-11: Matriculants to U.S. MD-Granting Medical Schools by Race/Ethnicity (Alone) and State of Legal Residence, 2023–2024*. <https://www.aamc.org/media/6041/download?attachment>
- Association of American Medical Colleges. (2025a). *MCAT exam scoring*. <https://students-residents.aamc.org/register-mcat-exam/publication-chapters/mcat-exam-scoring>
- Association of American Medical Colleges. (2025b). *What's on the MCAT exam? Exam overview*. <https://students-residents.aamc.org/whats-mcat-exam/publication-chapters/whats-mcat-exam>
- Association of American Medical Colleges. (2025c). *What's on the MCAT exam? PDF Outline*. <https://students-residents.aamc.org/prepare-mcat-exam/whats-mcat-exam-pdf-outline>
- Attewell, P. A., Lavin, D. E., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education*, 77(5), 886–924. <https://doi.org/10.1080/00221546.2006.11778948>
- Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, 29(2), 255–270. <https://doi.org/10.1016/j.econedurev.2009.09.002>
- Barnett, E., & Reddy, V. (2017). College placement strategies: Evolving considerations and practices. In K. L. McClarty, K. D. Mattern, & M. N. Gaertner (Eds.), *Preparing*

- students for college and careers: Theory, measurement, and educational practice* (pp. 82–93). Routledge.
- Bauer-Wolf, J. (2023, February 7). NCAA permanently ends SAT, ACT eligibility requirement for Division I, II student-athletes. *Higher Ed Dive*. <https://www.highereddive.com/news/ncaa-permanently-ends-sat-act-eligibility-requirement-division-i-ii/642117/>
- Beatty, A., Greenwood, M. R. C., & Linn, R. L. (Eds.). (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions*. National Academies Press.
- Belasco, A. S., Rosinger, K. O., & Hearn, J. C. (2015). The test-optional movement at America's selective liberal arts colleges: A boon for equity or something else? *Educational Evaluation and Policy Analysis*, 37(2), 206–223. <https://doi.org/10.3102/0162373714537350>
- Bennett, C. T. (2022). Untested admissions: Examining changes in application behaviors and student demographics under test-optional policies. *American Educational Research Journal*, 59(1), 180–216. <https://doi.org/10.3102/00028312211003526>
- Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science*, 20(7), 822–830. <https://doi.org/10.1111/j.1467-9280.2009.02368.x>
- Bingham, R. O., Chanatry, J., Baffi-Dugan, C., Childress, B., & Maxwell, S. A. (2012). Preparing for the new MCAT: The perspective of advisors. *Peer Review*, 14(4), 25–29.
- Black Lives Matter: About. (2024). *Black lives matter*. <http://blacklivesmatter.com/about/>.
- Boost, I. (2019, March 21). What the scammers got right about college admissions. *The Atlantic*. <https://www.theatlantic.com/education/archive/2019/03/real-college-bribery-scandal-whats-legal/585298/>
- Boshara, R., Emmons, W. R., & Noeth, B. J. (2015). The demographics of wealth—how age, education and race separate thrivers from strugglers in today's economy. *Essay No. 2: The Role of Education*. . Center for Household Financial Stability, Federal Reserve Bank of St. Louis. <https://www.stlouisfed.org/~media/Files/PDFs/HFS/essays/HFS-Essay-2-2015-Education-and-Wealth.pdf>
- Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton University Press.
- Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record*, 113, 2309–2344. <http://dx.doi.org/10.1177/0161468111111301101>
- Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic outcomes? *Educational Measurement: Issues and Practice*, 35(4), 21–24. <https://doi.org/10.1111/emip.12130>
- Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT® I: Reasoning Test* (College Board Research Report No. 00-1). ETS. <https://doi.org/10.1002/j.2333-8504.2000.tb01824.x>

- Burton, N. W., & Wang, M. M. (2014). *Predicting long-term success in graduate school: A collaborative validity study*. In C. Wendler, B. Bridgeman, & C. Ezzo, C. (Eds.), *The research foundation for the GRE revised General Test: A compendium of studies* (pp. 5.6.1–5.65). ETS. https://www.ets.org/s/research/pdf/gre_compendium.pdf
- Busche, K., Elks, M. L., Hanson, J. T., Jackson-Williams, L., Manuel, R. S., Parsons, W. L., Wofsy, D., & Yuan, K. (2020). The validity of scores from the new MCAT exam in predicting student performance: Results from a multisite study. *Academic Medicine*, 95(3), 387–395. <https://doi.org/10.1097/acm.0000000000002942>
- Camara, W. J. (2009). College admission testing: Myths and realities in an age of admissions hype. In R. Phelps (Ed.), *Correcting fallacies about educational and psychological testing* (pp. 147–180). American Psychological Association.
- Camara, W. J. (2018). *Illustrating differences among students with the same HSGPA of 4.0 or above* [Unpublished manuscript]. ACT.
- Camara, W. J. (2020). Never let a crisis go to waste: Large scale assessment and the response to COVID-19. *Educational Measurement: Issues and Practice*, 33(3), 10–18. <https://doi.org/10.1111/emip.12358>
- Camara, W. J. (2023). Admissions testing impact on access and alternative opinions. In K. Geisinger (Ed.), *College admissions testing* (pp. 10–31). Buros Institute.
- Camara, W. J. (2024). Admission testing in higher education: Changing landscape and outcomes from test-optional policies. *Educational Measurement: Issues and Practice*, 43, 104–111.
- Camara, W. J., & Davis, L. (2022). Fairness concerns resulting from innovations and applications of technology to assessment. In J. Jonson & K. Geisinger (Eds.), *Fairness issues and solutions in educational and psychological testing: Implications for researchers, practitioners, policy makers, and the public* (pp. 153–187). American Educational Research Association.
- Camara, W. J., Kimmel, E., Scheuneman, J., & Sawtell, E. (2003). Whose grades are inflated? (*CollegeBoardResearchReport2003–4*). CollegeBoard. http://professionals.collegeboard.com/profdownload/pdf/04843cbreport20034_31757.pdf
- Camara, W. J., & Mattern, K. (2022). Inflection point: The role of testing in admissions decisions in a post pandemic environment. *Educational Measurement: Issues and Practice*, 41, 10–15. <https://onlinelibrary.wiley.com/doi/epdf/10.1111/emip.12493>
- Camara, W. J., Mattern, K., Croft, M., Vispoel, S., & Nichols, P. (2019). A validity argument in support of the use of college admission test scores for federal accountability. *Educational Measurement: Issues and Practice*, 38(4), 12–26. <https://doi.org/10.1111/emip.12293>
- Camara, W. J., & Westrick, P. (2017). *USA perspectives on admissions and admissions testing*. ACT. <https://www.act.org/content/dam/act/unsecured/documents/R1653-usa-admission-perspectives-2017-09.pdf>
- Chen, X., & Simone, S. (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (Report No. 2016-405). U.S. Department of Education, Institute for Education Sciences, National Center for Education Statistics. <https://nces.ed.gov/pubs2016/2016405.pdf>

- Chetty, R., Deming, D. J., & Friedman, J. N. (2023). *Diversifying society's leaders? The determinants and causal effects of admission to highly selective private colleges* [Working paper]. National Bureau of Economic Research. <https://www.nber.org/papers/w31492>
- Clinedinst, M. (2019). *2019 state of college admission*. National Association for College Admission Counseling. https://nacacnet.org/wp-content/uploads/2022/10/soca2019_all.pdf
- College Board. (2016, May 9). *Concordance tables for new SAT scores*. <https://satsuite.collegeboard.org/media/pdf/higher-ed-brief-sat-concordance.pdf>
- College Board. (2017a). *SAT suite of assessments technical manual: Characteristics of the SAT*. <https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual.pdf>
- College Board. (2017b). *SAT suite of assessments technical manual appendixes*. <https://collegereadiness.collegeboard.org/pdf/sat-suite-assessments-technical-manual-appendix-pt-1.pdf>
- College Board. (2017c). *Next-generation ACCUPLACER: Test specifications (version 2.0)*. <https://accuplacer.collegeboard.org/accuplacer/pdf/next-generation-test-specifications-manual.pdf>
- College Board. (2018). *ACCUPLACER: Program manual*. <https://accuplacer.collegeboard.org/accuplacer/pdf/accuplacer-program-manual.pdf>
- College Board. (2019a). *SAT suite of assessments*. <https://reports.collegeboard.org/sat-suite-program-results/class-2019-results>
- College Board. (2019b). *ACCUPLACER suite of assessments technical manual*.
- College Board. (2024a, August). *Assessment framework for the digital SAT Suite (Version 3.01)*. <https://satsuite.collegeboard.org/media/pdf/assessment-framework-for-digital-sat-suite.pdf>
- College Board (2024b, March 12). *Digital SAT launches across the country, completing the transition to digital and providing a simpler testing experience for students and educators*. <https://newsroom.collegeboard.org/digital-sat-launches-across-country-completing-transition-digital-and-providing-simpler-testing#:~:text=03/12/2024-,Digital%20SAT%20Launches%20Across%20the%20Country%2C%20Completing%20the%20Transition%20to,Board's%20new%20digital%20testing%20app>
- College Board. (2024c). *Digital SAT suite of assessments: Technical manual*. <https://research.collegeboard.org/media/pdf/Digital%20SAT%20Suite%20of%20Assessments%20Technical%20Manual-FINAL.pdf>
- College Board & ACT. (2018). *Guide to the 2018 ACT/SAT concordance*. <https://www.act.org/content/dam/act/unsecured/documents/ACT-SAT-Concordance-Information.pdf>
- Cruce, T. M., & Sanchez, E. I. (2023, April 12–15). *Score-sending to test-optional colleges and first-year college success* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States.
- Dallessandro, S. P., Anthony, L. C., & Reese, L. M. (2014). *LSAT performance with regional, gender, and racial/ethnic breakdowns: 2007–2008 through 2013–2014 testing years* (LSAT Technical Report No. TR-14-02). Law School Admission Council.

- Davenport, E. K., Howard, B., & Harrington Weston, S. (2018). An examination of the effect of the University of Michigan cases on the complexion of higher education. *Alabama Journal of Educational Leadership*, 5, 29–38. <https://files.eric.ed.gov/fulltext/EJ1194727.pdf>
- Davies, M. (2011). Introduction to the special issue on critical thinking in higher education. *Higher Education Research & Development*, 30(3), 255–260. <https://doi.org/10.1080/07294360.2011.562145>
- de Brey, C., Zhang, A., & Duffy, S. (2022). *Digest of education statistics, 2020* (NCES 2022-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/programs/digest/d20/tables/dt20_305.40.asp
- Donlon, T. (1984). *College Board technical handbook for the Scholastic Aptitude Test and achievement tests*. College Entrance Examination Board Publishing.
- Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: A meta-analysis of the published research. *Academic Medicine*, 82, 100–106. <https://doi.org/10.1097/01.acm.0000249878.25186.b7>
- Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT assessment and recentered SAT I sum scores. *College and University*, 73(2), 24–32.
- Dudley, R. (2016, December 12). *College Board faces rocky path after CEO pushes new vision for SAT*. Reuters Investigates. <https://www.reuters.com/investigates/special-report/college-sat-coleman/>.
- Dunleavy, D. M., Kroopnick, M. H., Dowd, K. W., Searcy, C. A., & Zhao, X. (2013). The predictive validity of the MCAT exam in relation to academic performance through medical school: A national cohort study of 2001–2004 matriculates. *Academic Medicine*, 88(5), 666–671. <https://doi.org/10.1097/acm.0b013e3182864299>
- The Economist. (2014, April 5). *Is college worth it?* <http://www.economist.com/news/united-states/21600131-too-many-degrees-are-waste-money-return-higher-education-would-be-much-better>
- Education Commission of the States. (2022, December). *50-State comparisons: States' school accountability*. <https://www.ecs.org/50-state-comparison-states-school-accountability-systems/>
- Education Week. (2020, April 3). *State-by-state map of school-building closure*. <https://www.edweek.org/ew/section/multimedia/map-coronavirus-and-school-closures.html>
- Edwards, A., Goyer, P., Howell, J., Hurwitz, M., Imlay, S., & Perfetto, G. (2023). *College outcomes following pandemic-induced changes in college admissions policies*. College Board. https://research.collegeboard.org/media/pdf/ARC_Research_Brief_2_04032023.pdf
- ETS. (2017). *GRE to GMAT comparison tool for business schools*. <https://www.ets.org/content/dam/ets-org/pdfs/gre/gre-bschool-comparison-tool.xlsx>
- ETS. (2020a). *Taking the GRE general test for law school*. https://www.ets.org/gre/revised_general/about/law/

- ETS. (2020b). *The GRE® General Test*. <https://www.ets.org/gre/institutions/about/general?WT.ac=rx01>
- ETS. (2022). *GRE General Test preparation guide*. Retrieved from <https://www.ets.org/pdfs/gre/gre-guide-to-the-use-of-scores.pdf>
- ETS. (2023a). *GRE comparison tool for law schools*. <https://www.ets.org/gre/score-users/law-schools/comparison-tool.html>
- ETS. (2023b). *Learn about GRE*. <https://www.ets.org/gre/test-takers/general-test/about.html>
- ETS. (2023c). *The GRE Tests*. <https://www.ets.org/gre/score-users/about/general-test/content-structure.html>
- ETS. (2024). *GRE guide to the use of scores*. <https://www.ets.org/pdfs/gre/gre-guide-to-the-use-of-scores.pdf>
- FairTest (2024). *Test optional and test free colleges*. <https://fairtest.org/test-optional-list/>
- Fein, P. (2015, June 18). *Finding a new compass*. Inside Higher Education. <https://www.insidehighered.com/news/2015/06/18/act-drops-popular-compass-placement-test-acknowledging-its-predictive-limits>
- Feldon, D. F., Litson, K., Cahoon, B., Feng, Z., Walker, A., & Tofel-Grehl, C. (2023). The predictive validity of the GRE across graduate outcomes: A meta-analysis of trends over time. *Journal of Higher Education* 95(1), 120–148. <https://doi.org/10.1080/00221546.2023.2187177>
- Felton, E. (2015, October 29). How the Common Core is transforming the SAT. *The Atlantic*. <https://www.theatlantic.com/education/archive/2015/10/the-common-core-sat/412960/>
- Fields, R., & Parsad, B. (2012). *Tests and cut scores used for student placement in postsecondary education: Fall 2011*. National Assessment Governing Board. <https://www.nagb.gov/content/dam/nagb/en/documents/what-we-do/preparedness-research/surveys/test-and-cut-scores-used-for-student-placement-in-postsecondary-education-fall-2011.pdf>
- Finn, B. (2015). *Measuring motivation in low-stakes assessments* (ETS Research Report No. RR-15-19). ETS. <https://onlinelibrary.wiley.com/doi/10.1002/ets2.12067>
- Fisher v. University of Texas et al., 579 U.S. 136 (2016). <https://www.oyez.org/cases/2015/14-981>
- Fuess, C. M. (1950). *College Board: Its first fifty years*. Columbia University Press.
- Furuta, J. (2017). Rationalization and student/school personhood in U.S. college admissions: The rise of test-optional policies, 1987 to 2015. *Sociology of Education*, 90(3), 236–254. <https://doi.org/10.1177/0038040717713583>
- Gaff, J., & Meachum, J. (2006). Learning goals in mission statements: Implications for educational leadership. *Liberal Education*, 92(1), 6–13. <https://files.eric.ed.gov/fulltext/EJ743272.pdf>
- Gneezy, U., List, J. A., Livingston, J. A., Sadoff, S., Qin, X., & Xu, Y. (2017). *Measuring success in education: The role of effort on the test itself* (NBER Working Paper 24004). National Bureau of Economic Research.

- Goldberg, J., & Smith, J. (2008). The effects of education on labor market outcomes. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 688–708). Routledge.
- Graduate Management Admission Council. (2011). *Next-Generation GMAT exam*. https://www.gmac.com/gmat/next-generation-gmat-information-center/~media/files/gmac/gmat/2011_nextgenerationgmat_exam_v3.pdf
- Graduate Management Admission Council. (2018). *GMAT exam user guide for institutions*. https://www.gmac.com/-/media/files/gmac/gmat/gmat_institutionuserguide_fnl.pdf
- Graduate Management Admission Council. (2020). *GMAC history*. <https://www.gmac.com/why-gmac/the-gmac-advantage/gmac-history>
- Graduate Management Admission Council. (2023). *Exam structure, content & features*. <https://www.gmac.com/gmat-other-assessments/about-the-gmat-focus-edition/exam-overview>
- Graduate Management Admission Council. (2024a). *The GMAT Exam: Looking forward*. <https://www.gmac.com/gmat-other-assessments/about-the-gmat-focus-edition/gmat-focus-edition>
- Graduate Management Admission Council (2024b). *Understanding your score*. <https://www.mba.com/exams/gmat-exam/scores/understanding-your-score>
- Graduate Management Admission Council. (2025). *Scores*. <https://www.mba.com/exams/gmat-exam/scores>
- Gratz v. Bollinger et al., 539 U.S. 244 (2003). <https://www.oyez.org/cases/2002/02-516>
- Griffith, R. L., Wolfeld, L., Armon, B. K., Rios, J., & Liu, O. L. (2016). *Assessing intercultural competence in higher education: Existing research and future directions* (ETS Research Report No. RR-16-25). ETS. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12112>
- Grutter v. Bollinger et al., 539 U.S. 306 (2003). <https://www.oyez.org/cases/2002/02-241>
- Han, C. (2022, October 18). *GRE to GMAT conversion: Why you can't compare scores*. MBA.com. <https://www.mba.com/exams-and-exam-prep/gmat-exam/gmat-gre-why-you-cant-compare-scores>
- Han, K. T. (2025, October 17). *GMAT Focus psychometric updates* [Seminar presentation]. GMAT Test Preparation Summit, Reston, VA, United States.
- Hartocollis, A., & Saul, S. (2024, April 11). Harvard and Caltech will require test scores. *New York Times*. <https://www.nytimes.com/2024/04/11/us/harvard-test-scores-admissions.html>
- Hawkins, D. (2019). Operation Varsity Blues: Where we are now. *Journal of College Admission*, 244, 15–17. <https://eds.p.ebscohost.com/eds/pdfviewer/pdfviewer?vid=0&sid=b909a86e-6c63-4afd-8b10-1966b2c524b8%40redis>
- Hiss, W., & Franks, V. (2014). *Defining promise: Optional standardized testing policies in American college and university admissions*. Lumina Foundation. <https://www.luminafoundation.org/files/resources/definingpromise.pdf>

- Howell, J. (2023, April 12–15). *Insights from the admissions research consortium* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States.
- Hoxby, C., & Avery, C. (2013). *The missing “one-offs”: The hidden supply of high-achieving, low-income students*. Brookings Papers on Economic Activity. https://www.brookings.edu/wp-content/uploads/2016/07/2013a_hoxby.pdf
- Hurwitz, M., & Lee, J. (2018). Grade inflation and the role of standardized testing. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades and the future of college admissions* (pp. 64–93). Johns Hopkins University Press.
- Hyman, J. (2017). ACT for all: The effect of mandatory college entrance exams on postsecondary attainment and choice. *Education Finance and Policy*, 12(3), 281–311. https://doi.org/10.1162/edfp_a_00206
- Inside Higher Ed. (2019). *2019 survey of college and university presidents*. <https://www.insidehighered.com/news/survey/2019-survey-college-and-university-presidents>
- Jacobsen, E. (2017). *A (mostly) brief history of the SAT and ACT tests*. Erik the Red. <https://erikthered.com/tutor/sat-act-history.html>
- Jankowski, N. A., Timmer, J. D., Kinzie, J., & Kuh, G. D. (2018, January). *Assessment that matters: Trending toward practices that document authentic student learning*. University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment. <https://files.eric.ed.gov/fulltext/ED590514.pdf>
- Jaschik, S. (2017, September 18). *At University of Texas at Austin, 10% has become 6%*. Inside Higher Ed. <https://www.insidehighered.com/admissions/article/2017/09/18/university-texas-austin-adjusts-downward-percent-plan-used-admissions>
- Jaschik, S. (2018, April 9). *GMAT loses 30 minutes*. Inside Higher Ed. <https://www.insidehighered.com/admissions/article/2018/04/09/gmat-facing-competition-gre-cuts-time-test-30-minutes>
- Jaschik, S. (2020, October 19). *ACT and SAT scores drop*. Inside Higher Ed. <https://www.insidehighered.com/admissions/article/2020/10/19/act-and-sat-scores-drop-2020?msclkid=994ac0f4a60211eca79677b819f8b0b0>
- Jencks, C., & Phillips, M. (Eds.). (1998). *The Black–White test score gap*. Brookings Institution Press.
- Julian, E. R. (2005). Validity of the Medical College Admission Test for predicting medical school performance. *Academic Medicine*, 80(10), 910–917. <https://doi.org/10.1097/00001888-200510000-00010>
- Julian, T. (2012). *Work–life earnings by field of degree and occupation for people with a bachelor’s degree: 2011* (Report No. ACSBR/11-04). U.S. Census Bureau. <https://www.census.gov/library/publications/2012/acs/acsbr11-04.html>
- Kaniuka, T., & Wynne, M. (2019). Exploring the relationship between the Collegiate Learning Assessment, student learning activities, and study behaviors: Implications for colleges and universities. *Education Research International*, 2019(1), 1–11. <https://doi.org/10.1155/2019/3648318>
- Kaplan. (2021, February 9). *Kaplan survey: Nearly 30 percent of colleges say the Black Lives Matter movement changed their admissions process* [Press release]. <https://>

www.kaptest.com/blog/press/2021/02/09/kaplan-survey-nearly-30-percent-of-colleges-say-the-black-lives-matter-movement-changed-their-admissions-process/

- Keiser, H. N., Sackett, P. R., Kuncel, N. R., & Brothen, T. (2016). Why women perform better in college than admission scores would predict: Exploring the roles of conscientiousness and course-taking patterns. *Journal of Applied Psychology, 101*(4), 569–581. <https://doi.org/10.1037/apl0000069>
- Kelly, R., & Morgan, J. (2024). LSAT performance with regional, gender, racial and ethnic, repeater, and disability breakdowns: 2018-2019 through 2022-2023 testing years (LSAT Technical Report Series TR 24-01). LSAC. <https://www.lsac.org/sites/default/files/research/TR-24-01.pdf>
- Kent, J. D., & McCarthy, M. T. (2016). *Holistic review in graduate admissions: A report from the Council of Graduate Schools*. Council of Graduate Schools.
- Klasik, D. (2013). The ACT of enrollment: The college enrollment effects of state-required college entrance exam testing. *Educational Researcher, 42*(3), 151–160. <https://doi.org/10.3102/0013189X12474065>
- Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (2009). *Test validity study (TVS) report*. Collegiate Learning Assessment.
- Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). *The validity of GRE® General Test scores for predicting academic performance at U.S. law schools* (ETS Research Report No. RR-18-26). ETS. <https://doi.org/10.1002/ets2.12213>
- Klieger, D. M., Cline, F. A., Holtzman, S. L., Minsky, J. L., & Lorenz, F. (2014). *New perspectives on the validity of the GRE® general test for predicting graduate school grades* (ETS Research Report No. RR-14-26). ETS. <http://dx.doi.org/10.1002/ets2.12026>
- Knox, L. (2023, October, 19). No more “logic games” on the LSAT. *Inside Higher Ed*. <https://www.insidehighered.com/news/quick-takes/2023/10/19/lsat-drops-analytical-reasoning-section-after-lawsuit>
- Knox, L. (2024). Introducing a shorter, ‘more flexible’ ACT. *Inside Higher Ed*. <https://www.insidehighered.com/news/admissions/traditional-age/2024/07/22/act-announces-shorter-cheaper-science-optional-exam>
- Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. (2008). *Validity of the SAT for predicting first-year college grade point average* (Report No. 2008-5). College Board. <https://files.eric.ed.gov/fulltext/ED563202.pdf>
- Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning & Education, 6*(1), 51–68. <https://doi.org/10.5465/amle.2007.24401702>
- Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students’ success. *Science, 315*(5815), 1080–1081. <https://doi.org/10.1126/science.1136618>

- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127(1), 162–181. <https://doi.org/10.1037/0033-2909.127.1.162>
- Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, 70, 340–352. <https://doi.org/10.1177/0013164409344508>
- Laird, R. (2005). What is it we think we are trying to fix and how should we fix it? A view from the admissions office. In W. J. Camara & E. W. Kimmel (Eds.), *Choosing students: Higher education admissions tools for the 21st century* (pp. 13–32). Routledge.
- LaPrade, A., Mertens, J., Moore, T., & Wright, A. (2019). *The enterprise guide to closing skills gap, strategies for building and maintaining a skilled workforce*. Institute for Business Value. <https://www.ibm.com/thought-leadership/institute-business-value/report/closing-skills-gap>
- Lawrence, I., Rigol, G. W., Van Essen, T., & Jackson, C. A. (2002). *A historical perspective on the SAT 1926-2001* (Report No. 2002-7). College Board.
- Law School Admission Council. (2019). *How our test contributes to comprehensive and fair law school admissions*. <https://www.lsac.org/blog/how-our-test-contributes-comprehensive-and-fair-law-school-admissions>.
- Law School Admission Council. (2020). *LSAT trends: Total LSATs administered by admin & year*. <https://report.lsac.org/TestTakers.aspx?Format=PDF>
- Law School Admission Council. (2022). *2021–22 interpretive guide for LSAT® score users*. LSAC.
- Law School Admission Council. (2023a). *Summary of 2017, 2018, and 2019 LSAT correlation study results*. (2nd ed.) <https://www.lsac.org/data-research/research/summary-2017-2018-and-2019-lsat-correlation-study-results>
- Law School Admission Council. (2023b). *What to expect starting with the August 2024 LSAT*. <https://www.lsac.org/blog/what-to-expect-starting-with-august-2024-lsat>
- Law School Admission Council. (2024, December 18.) *Admission trends: Applicants, admitted applicants, matriculants, & applications*. <https://report.lsac.org/View.aspx?Report=AdmissionTrendsApplicantsAdmitApps&Format=PDF>
- Lawyers for Civil Rights v. President and Fellows of Harvard College. Complaint before U.S. Department of Education (2023, July 3).
- Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.
- Lewin, T. (2009, December 9). A new look for graduate entrance test. *The New York Times*. <https://www.nytimes.com/2009/12/06/education/06gre.html>
- Li, D., Yi, Q., & Harris, D. (2017). *Evidence for paper and online ACT comparability: Spring 2014 and 2015 mode comparability studies* (Report No. 2017-1). ACT.
- Lindquist, E. F. (1959). The nature of the problem of improving scholarship and college entrance examinations. In *Proceedings of the Invitational Conference on Testing Problems* (pp. 104–113). ETS.

- Liu, O. L. (2017). Ten years after the Spellings Commission: From accountability to internal improvement. *Educational Measurement: Issues and Practice*, 36(2), 34–41. <http://dx.doi.org/10.1111/emip.12139>.
- Liu, O. L. (2023). Assessing student learning outcomes of higher education. In C. Baik & E. Kahu (Eds.), *Research handbook on the student experience in higher education* (pp. 283–295). Edward Elgar Publishing.
- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes assessment in higher education: Motivation matters. *Educational Researcher*, 41, 352–362. <https://psycnet.apa.org/doi/10.3102/0013189X12459679>
- Liu, O. L., Bridgeman, B., Gu, L. X., Xun, J., & Kong, N. (2015). Investigation of response change on the revised GRE. *Educational & Psychological Measurement*, 75(6), 1002–1020. <https://dx.doi.org/10.1177%2F0013164415573988>
- Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next generation assessment* (ETS Research Report No. RR-14-10). ETS. <https://doi.org/10.1002/ets2.12009>
- Liu, O. L., Kell, H. J., Liu, L., Ling, G., Wang, Y., Wylie, C., Sevak, A., Sherer, D., LeMahieu, P., & Knowles, T. (2023). *A new vision for skills-based assessment*. ETS. <https://www.ets.org/pdfs/rd/new-vision-skills-based-assessment.pdf>
- Liu, O. L., Kell, H., Williams, K., Ling, G. M., & Sanders, M. (2023). ETS 2025 skills taxonomy. *Chinese/English Journal of Educational Measurement and Evaluation*. 4(4), Article 2. <https://doi.org/10.59863/URZL9477>
- Liu, O. L., Klieger, D. M., Bochenek, J. L., Holtzman, S. L., & Xu, J. (2016). *An investigation of the use and predictive validity of scores from the GRE® revised General Test in a Singaporean university* (ETS Research Report No. RR-16-05). ETS. <https://doi.org/10.1002/ets2.12095>
- Liu, O. L., Liu, H., Roohr, K. C., & McCaffrey, D. F. (2016). Investigating college learning gain: Exploring a propensity score weighting approach. *Journal of Educational Measurement*, 53(3), 352–367. <https://doi.org/10.1111/jedm.12112>
- Liu, O. L., Ling, G., & Fishtein, D. (2021). Charting the skills for the future: Perspectives from employers and higher education institutions. *NCME Chinese/English Journal of Educational Measurement and Evaluation* 3(1), 24–32.
- Lucido, J. A. (2018). Understanding the test-optional movement. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades and the future of college admissions* (pp. 146–170). Johns Hopkins University Press.
- Magouirk, P., Freeman, M., Kajkawa, T., Karimi, H., & Kim, B. H. (2023, January 11). *Deadline update: first-year application trends through January 1*. Common App. https://s3.us-west-2.amazonaws.com/ca.research.publish/Deadline+Updates/DeadlineUpdate_010323.pdf
- Marco, G. L., Abdel-fattah, A. A., & Baron, P. A. (1992). *Methods used to establish score comparability on the enhanced ACT assessment and the SAT* (Research Report No. 92-3). College Board. <https://doi.org/10.1002/J.2333-8504.1992.TB01456.X>
- Margolis, W. (2018, December 10). *LSAC announces technology collaboration with Microsoft*. Law School Admission Council. <https://www.lsac.org/about/news/lisac-announces-technology-collaboration-microsoft>

- Marini, J. P., Westrick, P. A., Young, L., Ng, H., & Shaw, E. J. (2023). *Digital SAT pilot predictive validity study—A first look*. College Board.
- Marini, J. P., Westrick, P. A., Young, L., Ng, H., Shmueli, D., & Shaw, E. J. (2019). *Differential validity and prediction of the SAT®: Examining first-year grades and retention to the second year*. College Board. <https://files.eric.ed.gov/fulltext/ED597325.pdf>
- Marion, S., & Domaleski, C. (2019). An argument in search of evidence: A response to Camara et al. (2019). *Educational Measurement: Issues and Practice*, 38(4), 27–28. <http://doi.org/10.1111/emip.12307>
- Mattern, K., & Allen, J. (2016). *More information, more informed decisions: Why test-optional policies do not benefit institutions or students*. ACT. https://www.act.org/content/dam/act/unsecured/documents/R1603_More-Information-More-Informed-Decisions_Web.pdf
- Mattern, K. D., & Packman, S. (2009). *Predictive validity of ACCUPLACER scores for course placement: A meta-analysis* (Research Report No. 2009-2). College Board. <https://files.eric.ed.gov/fulltext/ED561046.pdf>
- Mattern, K. D., & Patterson, B. F. (2014). *Synthesis of recent SAT validity findings: Trend data over time and cohorts* (College Board Research in Review 2014-1). College Board. <https://files.eric.ed.gov/fulltext/ED556462.pdf>
- Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. (2008). *Differential validity and prediction of the SAT* (Research Report No. 2008-4). College Board. <https://files.eric.ed.gov/fulltext/ED562614.pdf>
- Mattern, K., Radunzel, J., Bertling, M., & Ho, A. D. (2018). How should colleges treat multiple admissions test scores? *Educational Measurement: Issues and Practice*, 37(3), 11–23. <https://doi.org/10.1111/emip.12199>
- Mattern, K., Sanchez, E., & Ndum, E. (2017). Why do achievement measures underpredict female academic performance? *Educational Measurement: Issues and Practice*, 36(1), 47–57. <https://psycnet.apa.org/doi/10.1111/emip.12138>
- Mattern, K., & Steedle, J. (2020). *Students who take the ACT test sections in a different order earn similar scores* [ACT Research Report 2020-1]. ACT. <https://www.act.org/content/dam/act/unsecured/documents/R1822-section-order-10-2020.pdf>
- McGaghie, W. (2002). Assessing readiness for medical education: Evolution of the Medical College Admission Test. *Journal of the American Medical Association*, 288(9), 1085–1090. <https://doi.org/10.1001/jama.288.9.1085>
- McKinsey Global Institute. (2017, December). *Jobs lost, jobs gained: Workforce transitions in a time of automation*. McKinsey & Company. <https://www.mckinsey.com/~media/mckinsey/featured%20insights/Future%20of%20Organizations/What%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/MGI-Jobs-Lost-Jobs-Gained-Report-December-6-2017.ashx>
- Michel, R. S., Belur, V., Naemi, B., & Kell, H. J. (2019). *Graduate admissions practices: A targeted review of the literature* (ETS Research Report No. RR-19-33). ETS. <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ets2.12271>
- Monroe, A., Quinn, E., Samuelson, W., Dunleavy, D. M., & Dowd, K. W. (2013). An overview of the medical school admission process and use of applicant data in

- decision making: What has changed since the 1980s? *Academic Medicine*, 88(5), 672–281. <https://doi.org/10.1097/acm.0b013e31828bf252>
- National Association of Colleges and Employers. (2018). *Job outlook 2019*. <https://www.odu.edu/content/dam/odu/offices/cmc/docs/nace/2019-nace-job-outlook-survey.pdf>
- National Center for Education Statistics. (2018). *Digest of education statistics, Table 302.10. Recent high school completers and their enrollment in college, by sex and level of institution: 1960 through 2017*. U.S. Department of Education, Institute of Education Sciences. https://nces.ed.gov/programs/digest/d18/tables/dt18_302.10.asp
- National Council on Measurement in Education. (2019). *National Council on Measurement in Education position statement on the use of college admissions test scores as academic indicators in state accountability systems*. <https://doi.org/10.1111/emip.12309>
- National Institute for Learning Outcomes Assessment. (2019, September). *Student learning outcomes statements*. <https://www.learningoutcomesassessment.org/ourwork/transparency-framework/components/student-learning-outcomes-statements/>
- National Student Clearinghouse Research Center. (2019). *Snapshot report: First-year persistence and retention for fall 2017 cohort*. <https://nscresearchcenter.org/wp-content/uploads/SnapshotReport35.pdf>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw–Hill.
- Office of the President of University of California. (2020, May 21). *Executive summary: College entrance exam use in University of California undergraduate admissions*. <https://regents.universityofcalifornia.edu/regmeet/may20/b4.pdf>
- Oh, I. S., Schmidt, F. L., Shaffer, J. A., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning & Education*, 7(4), 563–570. <https://doi.org/10.5465/amle.2008.35882196>
- Olson, E. (2017, August 10). More law schools begin accepting GRE test results. *New York Times*. <https://www.nytimes.com/2017/08/10/business/dealbook/law-school-gre.html>
- Organisation for Economic Co-operation and Development. (2013). *Assessment of higher education learning outcomes: Feasibility study report: Vol. 3. Further insights*. <http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume3.pdf>
- Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Research Council, The National Academies Press. <https://www.nap.edu/read/13398/chapter/1>
- Posselt, J. R. (2016). *Inside graduate admissions: Merit, diversity, and faculty gatekeeping*. Harvard University Press.
- Powers, D. E. (2004). Validity of Graduate Record Examinations (GRE) General Test scores for admissions to colleges of veterinary medicine. *Journal of Applied Psychology*, 89, 209–219. <https://doi.org/10.1037/0021-9010.89.2.208>

- Powers, D. E., & Fowles, M. E. (2000). *Likely impact of the GRE writing assessment on graduate admissions decisions* (ETS Research Report No. RR-00-16). ETS. <http://dx.doi.org/10.1002/j.2333-8504.2000.tb01839.x>
- Radunzel, J., & Mattern, K. (2020a). *Predicting students' chances of completing a degree: How does superscoring compare to other scoring methods when applicants retest?* (Working Paper 2020-01). ACT. <https://www.act.org/content/dam/act/unsecured/documents/R1825-degree-completion-superscore-2020-05.pdf>
- Radunzel, J., & Mattern, K. (2020b). *A case study: ACT section retest scores and superscores are predictive of first-term grades*. ACT. <https://files.eric.ed.gov/fulltext/ED606153.pdf>
- Radunzel, J., Mattern, K., Crouse, J., & Westrick, P. (2015). *Development and validation of a STEM benchmark based on the ACT STEM score*. ACT. <https://www.act.org/content/dam/act/unsecured/documents/2015-Tech-Brief-Development-and-Validation.pdf>
- Radunzel, J., & Noble, J. (2012). *Predicting long-term college success through degree completion using ACT Composite score, ACT Benchmarks, and high school grade point average* (Report No. 2012-5). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2012-5.pdf
- Radunzel, J., & Noble, J. (2013). *Differential effects on student demographic groups of using ACT college readiness assessment composite score, ACT benchmarks, and high school grade point average for predicting long-term college success through degree completion* (Report No. 2013-5). ACT.
- Rahman, N., & Trierweiler, T. J. (2017). *Analysis of differential prediction of law school performance by race/ethnicity based on 2011–2014 entering law school classes* (LSAT Technical Report No. TR-17-01). Law School Admission Council.
- Reese, L. M., & Cotter, R. A. (1994). *A compendium of LSAT and LSAC-sponsored item types 1948–1994* (LSAT Research Report No. RR-94-01). Law School Admission Council. <https://files.eric.ed.gov/fulltext/ED469242.pdf>
- Richer, A. D., & Binkley, C. (2019, March 12). *TV stars and coaches charged college bribery scheme*. Associated Press. <https://apnews.com/2450688f9e67435c8590e59a1b0e5b47>
- Rigol, G. (2003). *Admissions decision-making models: How US institutions of higher education select undergraduate students*. College Entrance Examination Board. <https://files.eric.ed.gov/fulltext/ED562589.pdf>
- Rigol, G. (2004). *Selection through individual review: A report on phase IV of the admissions models project*. College Entrance Examination Board. <https://files.eric.ed.gov/fulltext/ED562855.pdf>
- Rippetoe, R. (2021, June 20). *Winston & Strawn aids UC students in test score case*. Law360. <https://www.law360.com/articles/1394856/winston-strawn-aids-uc-students-in-test-score-case>
- Rodriguez, O., Bowden, B., Belfield, C., & Scott-Clayton, S. (2015). *Calculating the costs of remedial placement testing*. Community College Research Center. <https://ccrc.tc.columbia.edu/media/k2/attachments/calculating-cost-remedial-placement-analytics-2.pdf>

- Roohr, K. C., Graf, E. A., & Liu, O. L. (2014). *Assessing quantitative literacy in higher education: An overview of existing research and assessments with recommendations for next generation assessment* (ETS Research Report No. RR-14-22). ETS. <https://doi.org/10.1002/ets2.12024>
- Roohr, K., Olivera-Aguilar, M., Ling, G., & Rikoon, S. (2019). A multi-level modeling approach to investigating students' critical thinking at higher education institutions. *Assessment & Evaluation in Higher Education*, 44(2), 1–15. <https://doi.org/10.1080/02602938.2018.1556776>
- Rutschow, E. Z., & Mayer, A. K. (2018). *Early findings from a national survey of developmental education practices* [Research brief]. Center for the Analysis of Postsecondary Readiness. https://www.mdrc.org/sites/default/files/2018_CAPR_Descriptive_Study.pdf
- Sanchez, E. I. (2013). *Differential effects of using ACT college readiness assessment scores and high school GPA to predict first-year college GPA among racial/ethnic, gender, and income groups* (Report No. 2013-4). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2013-4.pdf
- Sawyer, R. (2010). *Usefulness of high school average and ACT scores in making college admission decisions* (Report No. 2010-2). ACT. https://www.act.org/content/dam/act/unsecured/documents/ACT_RR2010-2.pdf
- Sawyer, R. (2013). Beyond correlations: Usefulness of high school GPA and test scores in making college admissions decisions. *Applied Measurement in Education*, 26(2), 89–112. <https://doi.org/10.1080/08957347.2013.765433>
- Schemo, D. J. (2001, February 17). Head of U. of California seeks to end SAT use in admissions. *New York Times*. <https://www.nytimes.com/2001/02/17/us/head-of-u-of-california-seeks-to-end-sat-use-in-admissions.html>
- Schudson, M. S. (1972). Organizing the “meritocracy”: A history of the college entrance examination board. *Harvard Educational Review*, 42(1), 34–69. <https://doi.org/10.17763/haer.42.1.15772104r6108v01>
- Schwager, I. T., Hülshager, U. R., Bridgeman, B., & Lang, J. W. (2015). Graduate student selection: Graduate Record Examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European University. *International Journal of Selection and Assessment*, 23(1), 71–79. <https://doi.org/10.1111/ijsa.12096>
- Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success?* (CCRC Working Paper No. 41). Columbia University, Teachers College, Community College Research Center. <http://ccrc.tc.columbia.edu/media/k2/attachments/high-stakes-predict-success.pdf>
- Shapiro, D., Dundar, A., Huie, F., Wakhungu, P. K., Bhimdiwali, A., Nathan, A., & Youngsik, H. (2018). *Transfer and mobility: A national view of student movement in postsecondary institutions, fall 2011 cohort* (Signature Report No. 15). National Student Clearinghouse.
- Shaw, E. J., Marini, J. P., Beard, J., Shmueli, D., Young, L., & Ng, H. (2016). *The redesigned SAT® pilot predictive validity study: A first look* (Report No. 2016-1). College Board.

- Shelton, P. D. (1997). The LSAT: Good—but not that good. *Law School Services Report*, 97, 2–3.
- Shultz, M. M., & Zedeck, S. (2011). Predicting lawyer effectiveness: Broadening the basis for law school admission decisions. *Law & Social Inquiry*, 36(3), 620–661. <https://doi.org/10.1111/j.1747-4469.2011.01245.x>
- Smith v. Regents of University of California et al. (2020). <https://unicourt.com/case/ca-sca1-smith-v-regents-of-the-university-of-california-et-al-131183>
- Snyder, T. D., de Brey, C., & Dillow, S. A. (2019). *Digest of Education Statistics 2017* (NCES 2018-070; Table 219.10, p. 193). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. <https://nces.ed.gov/pubs2018/2018070.pdf>
- Soares, J. A. (2012). *SAT wars: The case for test-optional college admissions*. Teachers College.
- Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). *Assessing written communication in higher education: Review and recommendations for next generation assessment* (ETS Research Report No. RR-14-37). ETS. <https://doi.org/10.1002/ets2.12035>
- Standardized Testing Task Force. (2020, January). *Report of the UC Academic Council Standardized Testing Task Force (UC Task Force)*. https://senate.universityofcalifornia.edu/_files/committees/sttf/sttf-report.pdf
- Stilwell, L. A., Dalessandro, S. P., & Reese, L. M. (2011). *Predictive validity of the LSAT: A national summary of the 2009 and 2010 LSAT correlation studies* (LSAT Technical Report TR-11-02). Law School Admission Council. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.225.3802&rep=rep1&type=pdf>
- Stilwell, L. A., Diamond-Dalessandro, S. P., & Reese, L. M. (2007). *Predictive validity of the LSAT: A national summary of the 2005–2006 correlation studies* (LSAT Technical Report TR-07-02). Law School Admission Council.
- Stilwell, L. A., Diamond-Dalessandro, S. P., & Reese, L. M. (2009). *Predictive validity of the LSAT: A national summary of the 2007–2008 correlation studies* (LSAT Technical Report TR-09-03). Law School Admission Council.
- Students for Fair Admissions v. President and Fellows of Harvard College. (2023). Oyez. <https://www.oyez.org/cases/2022/20-1199>.
- Students for Fair Admissions v. University of North Carolina. (2023). Oyez. <https://www.oyez.org/cases/2022/21-707>.
- Sweitzer, K., Blaock, A., E., Sharma, D. B. (2018). The effect of going test-optional on diversity and admissions: A propensity score matching analysis. In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades and the future of college admissions* (pp. 288–307). Johns Hopkins University Press.
- Syverson, S. T., Franks, V. W., & Hiss, W. C. (2018). *Defining access: How test-optional works*. NACAC. <https://www.nacacnet.org/globalassets/documents/publications/research/defining-access-report-2018.pdf>
- Talento-Miller, E. (2017). *Differential validity and differential prediction of the GMAT Exam*. Graduate Management Admission Council. <https://www.gmac.com/-/media/files/gmac/research/validity-and-testing/rr-17-01-differential-validity-talento-miller-web-release-2.pdf>

- Torney-Purta, J., Cabrera, J. C., Roohr, K. C., Liu, O. L., & Rios, J. A. (2015). *Assessing civic competency and engagement in higher education: Research background, frameworks, and directions for next-generation assessment* (ETS Research Report No. RR-15-34). ETS. <http://dx.doi.org/10.1002/ets2.12081>
- Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes (AHELO): Feasibility study report: Vol 1. Design and implementation*. OECD. www.oecd.org/edu/skills-beyond-school/AHELOFSReportVolume1.pdf
- Trierweiler, T. J., & Rahman, N. (2017). *Analysis of differential prediction of law school performance by gender based on 2011–2014 entering law school classes* (LSAT Technical Report TR-17-02). Law School Admission Council. <https://www.lsac.org/data-research/research/analysis-differential-prediction-law-school-performance-gender-based-2011>
- University of California, Ethics, Compliance, and Audit Services. (2020, February). *System wide audit of undergraduate admissions—Phase 2* (Project No. P20A005). <https://www.universityofcalifornia.edu/sites/default/files/second-audit-report.pdf>
- U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. The Secretary of Education's Commission on the Future of Higher Ed. <https://files.eric.ed.gov/fulltext/ED493504.pdf>
- U.S. Department of Education. (2015). *Every Student Succeeds Act (ESSA)*. <https://www.ed.gov/essa?src=rn>
- Wagenaar, R. (Ed.). (2018). *Tuning–CALOHEE assessment reference frameworks for civil engineering teacher education history nursing physics*. University of Groningen. <https://www.calohee.eu/wp-content/uploads/2018/11/4.1-Assessment-Reference-Frameworks-for-Civil-Engineering-Teacher-Education-History-Nursing-and-Physics-FINAL-v2.pdf>
- Wendler, C., Bridgeman, B., & Ezzo, C. (2014). *The research foundation for the GRE revised General Test: A compendium of studies*. ETS. https://www.ets.org/s/research/pdf/gre_compendium.pdf
- Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment*, 20(1), 23–45. <http://dx.doi.org/10.1080/10627197.2015.997614>
- Westrick, P. A., Marini, J. P., Shmueli, D., Young, L., Shaw, E. J., & Ng, H. (2020). *Validity of SAT® for predicting first-semester, domain-specific grades*. College Board. <https://files.eric.ed.gov/fulltext/ED603085.pdf>
- Westrick, P. A., Marini, J. P., Young, L., Ng, H., Shmueli, D., & Shaw, E. J. (2019). *Validity of the SAT for predicting first-year grades and retention to the second year*. College Board. <https://satsuite.collegeboard.org/media/pdf/national-sat-validity-study.pdf>
- Westrick, P. A., Schmidt, F. L., Le, H., Robbins, S. B., & Radunzel, J. M. (2021). The road to retention passes through first-year academic performance: A meta-analytic path analysis of academic performance and persistence. *Educational Assessment*, 26(1), 35–51. <https://doi.org/10.1080/10627197.2020.1848423>

- Whitney, D. R. (1989). Education admissions and placement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 515–526). Macmillan Publishing Co., American Council on Education.
- Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. College Entrance Examination Board.
- Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Lawrence Erlbaum Associates.
- Wise, S. L., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. http://dx.doi.org/10.1207/s15326977ea1001_1
- Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–301). Routledge.
- Young, J. W., Klieger, D., Bochenek, J., Li, C., & Cline, F. (2014). *The validity of scores from the GRE® revised general test for forecasting performance in business schools: Phase one* (ETS Research Report No. RR-14-17). ETS. <https://files.eric.ed.gov/fulltext/EJ1109333.pdf>
- Zamudio-Suarez, F. (2022, July 12). *Is remedial ed holding students of color back?* Race on Campus. <https://www.chronicle.com/newsletter/race-on-campus/2022-07-12>
- Zhou, E., & Gao, J. (2021). *Graduate enrollment and degrees: 2010 to 2020*. Council of Graduate Schools. https://cgsnet.org/wp-content/uploads/2022/02/CGS_GED20_Report_final_v2-2.pdf
- Zimmermann, J., von Davier, A., Buhmann, J. M., & Heinemann, H. R. (2017). Validity of GRE General Test scores and TOEFL scores for graduate admission to a technical university in western Europe. *European Journal of Engineering Education*, 43, 144–165. <https://doi.org/10.1080/03043797.2017.1343277>
- Zlatkin-Troitschanskais, O., Toepper, M., Pant, H. A., Lautenbach, C., & Kuhn, C. (Eds.). (2018). *Methodology of educational measurement and assessment: Assessment of learning outcomes in higher education, cross-national comparisons and perspectives*. Springer.
- Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. Routledge.
- Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647–679). Praeger Publishers.
- Zwick, R. (2013). *Disentangling the role of high school grades, SAT® scores, and SES in predicting college achievement*. ETS Research Report No. RR-13-09.
- Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions*. Harvard University Press.
- Zwick, R. (2018). How do percent plans and other test-optional admissions programs affect the academic performance and diversity of the entering class? In J. Buckley, L. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades and the future of college admissions* (pp. 239–259). Johns Hopkins University Press.

NOTES

1. Data were obtained through the National Center for Education Statistics Integrated Postsecondary Education Data System data center at <https://nces.ed.gov/ipeds/datacenter/reportmain.aspx>. Enrollment of all students at test-optional colleges represented about 5% of all freshmen nationally in 2018, but since about 70%–80% of students at such colleges submit test scores, the estimate is that students admitted without scores represent 1%–1.25% of freshmen.
2. See page 1194 in this chapter for a discussion of *Smith v. Regents of University of California*, 2020.
3. See Rigol (2004) for an example of an admissions index chart, which illustrates how admissions decisions could be made in a compensatory fashion allowing higher performance in either test scores or grades to serve a compensatory role for lower test scores or grades.
4. Plaintiffs failed to gain acceptance by automatic admission where the top 10% of students in each Texas high school's graduating class were accepted to the University of Texas at Austin, irrespective of race, and were denied admission during a second review for students not automatically admitted. In 2009, legislation limited automatic admissions to 75% of the university's enrollment capacity, which has lowered the automatic admission to the top 6% at each high school (Jaschik, 2017).
5. Of the 58 items, 38 prohibit calculator use.
6. The last official concordance study sanctioned by both testing programs and tables can be found at ACT (2018).
7. The ACT was administered exclusively on computer for international test takers, and available by computer for some state and district school-based administrations prior to 2024.
8. Open admission is an unselective and noncompetitive college admissions process in the United States, where the only requirement is a high school diploma or a General Educational Development certificate. The number of open-admission institutions was obtained through the National Center for Education Statistics Integrated Postsecondary Education Data System data center at <https://nces.ed.gov/ipeds/datacenter/reportmain.aspx>. Sorting variables for the 2018–2019 year included admissions and test scores, open admission policy (yes); and the following institutional characteristics: level of institution (1 four or more years); sector of institution (1 public, 4 years or above; and 2 private, not-for-profit, 4 years or above); degree-granting status (1 degree granting); and undergraduate offering (1 undergraduate degree or certificate offering). Total enrollment was about 2 million at these open-admission institutions. Some of the institutions were branch campuses of institutions that had selective admissions at their main campus.

9. Includes individuals ages 16–24 who graduated high school or completed a General Educational Development certificate or other equivalency program. See https://nces.ed.gov/programs/coe/indicator_cpa.asp.
10. The ACT and SAT report degree aspirations separately and with slightly different categories. The percentages were calculated by a weighted average from ACT and SAT data, which may result in double counts of students who completed both tests.
11. The National Collegiate Athletics Association suspended and has now eliminated its test score requirement for Division I or II athletics (Bauer-Wolf, 2023).
12. After July, 2024, the name transitioned back to the *GMAT Exam* (Graduate Management Admission Council, 2024a).
13. Two fixed-length paper forms called companion tests had been developed for each of the five adaptive tests.

Assessment for Licensing and Certification

Melissa J. Margolis
NBME

Rebecca S. Lipner
American Board of Internal Medicine

Chad W. Buckendahl
ACS Ventures

Regulation, credentialing, licensure, certification: These terms refer to aspects of occupational oversight that may occur at the federal level, the state level, the intra-professional level, or a combination of all three. Regardless of the responsible entity or entities, this oversight serves the primary purpose of ensuring that practitioners demonstrate a specific level of competence to protect themselves, the profession, and, most important, the public from harm.

Regulation tends to refer more generally to higher level oversight of different aspects of a profession. In essence, professional regulation can refer to “any activity that is intended to promote and protect the public interest by reducing, suppressing, mitigating or eliminating harms or potential harms stemming from the practice of a profession” (Balthazard, 2017). In some domains, professional regulation is, relatively speaking, straightforward: Professional engineering practice, for example, is governed by individual states that have specified educational, internship, and examination licensure requirements. The work of the state licensing boards is supported by a regulatory body, the National Society of Professional Engineers, which promotes strong licensure laws intended to protect “public health, safety, and welfare” (2023, Our Values section). Other domains, such as medicine, have a multifaceted and complex regulatory process involving numerous required or elected entities that have interconnected levels of oversight throughout a given physician’s professional career (White, 2014).

Credentialing is a term that can refer to the process of verifying eligibility for participating in a profession. Though such verification may occur at multiple points along the continuum of education and training (e.g., verifying completion of the educational requirements for taking a required examination, verifying successful completion of required examinations), the ultimate credentialing activity requires verification of *all* credentials that are necessary for performing the required job. Examples of credentials include licensing and certification, as well as other types of designations such as professional certificates, badges, or microcredentials. For the purposes of this chapter, the examples focus primarily on professional licensing and certification rather than on other types of credentials that have more recently gained traction in the professional arena. When relevant to the discussion, these other designations will be considered.

Licensure and *certification* typically refer to processes that are intended to provide the public with evidence that certain professional standards have been met; through these two processes, members of a profession are granted certain privileges that are associated with that profession (Raymond, 2015). A historic, yet somewhat tenuous, distinction between the two credentials relates to whether obtaining the credential is mandatory or voluntary. With licensing, a governmental agency grants permission to engage in an occupation after the specified requirements have been met; these requirements usually relate to educational activities, experience working in a supervised setting, and one or more examinations that a candidate must complete to demonstrate minimum competence with respect to the knowledge, skills, and abilities that are required for practice. For those occupations that require it, licensing is mandatory; without a license, one is not permitted to engage in the occupation even if all other requirements have been

met. Alternatively, certification typically is *not* a requirement for occupational practice. Instead, it traditionally has been viewed as an activity that is designed to demonstrate competence for entering an advanced level of practice; as such, it is more a reflection of professional oversight rather than the governmental oversight that applies to licensing. To further complicate matters, certification in business and industry may be used to indicate mastery of knowledge and skills associated with products or services, rather than as a gateway into a field (e.g., technology hardware, software, medical devices). Though the historical approach to distinguishing between licensure and certification based on the mandatory versus voluntary nature of the process has been popular and in many cases is accurate, there are exceptions to this general rule. In some professions, such as teaching, certification is mandatory and acts as a *de facto* license (Raymond, 2015).¹ There are professions in which both credentialing practices are relevant, and the distinction between them is clear. In the licensing of allopathic physicians, for example, licensure is governed by the state medical board; candidates must meet the educational, training, and assessment requirements for licensure and then apply to the state for the undifferentiated license to practice medicine. Should a licensed physician desire specialty certification, they must complete additional required training in that discipline and then pass one or more certification examinations administered by the specialty board; passing the examination(s) allows that diplomate to use the associated specialty designation (e.g., a board-certified plastic surgeon). In other professions, the distinction between licensing and certification is less clear. The medical example leads to consideration of an additional distinction that often is used to differentiate the two credentials: Licensing tends to be conceptualized as more of a hurdle to enter a given profession, whereas certification tends to reflect a more advanced level of practice (Kane et al., 1999). There are additional complexities associated with variability in credentialing requirements within a given field. In psychology, for example, health service (clinical, counseling, and school) psychologists typically require a license to practice, whereas the same credential may not be required in other applied psychological domains (DeMers et al., 2014; Elchert, 2016). With few exceptions, because licensing in the United States is a function of individual states/territories, there is variability across jurisdictions regarding which professions require licenses and what is required to maintain them.

Despite differences in terminology or specific requirements, the overarching goal of both credentialing processes is the same: to ensure that the practitioner has attained at least a minimum level of competence in the profession to protect the welfare of the public. As we have mentioned, professional credentialing organizations typically specify at least three types of requirements for receiving the credential: educational, experiential, and examination. This is not universal, however, and many certifications do not have education or experience as eligibility requirements to sit for an examination. The specific educational and experiential requirements for licensure and certification vary greatly across professions and jurisdictions; considering the third requirement—completion of one or more examinations—allows for focusing on the similarities in requirements that exist across these high-stakes contexts.

Each year, hundreds of thousands of test takers complete examinations that are designed to provide information about the extent to which they have met certain professional standards (Raymond, 2015). Because licensing and certification examinations are intended to inform decisions about a candidate's ability to perform effectively at the relevant level of practice (Kane et al., 1999), developing these tests focuses specifically on identifying the necessary requirements for professional practice at the level of interest (e.g., entry level for licensing, a more advanced level for certification); this allows for greater confidence in the link between performance on the assessment and effective job performance. Additional important considerations with respect to these examinations relate to topics such as legal issues associated with using tests to make decisions about the ability to enter practice (or practice using a specific title), the importance of methodological rigor throughout the item and test development process, the extent to which relevant stakeholders are included in the process, elements of standard setting that are unique to or warrant special attention in credentialing contexts, and other factors that are relevant to settings in which results of the assessment process have significant professional implications. By outlining these critical areas, the chapter aims to support the development of more effective, equitable, and evidence-based assessment practices in high-stakes professional contexts.

As should be clear from the preceding pages, we present a broad review of considerations that are applicable to professional licensing and certification contexts. Other chapters in this volume delve into many of these issues in substantial detail, and the interested reader is referred to these chapters for an in-depth discussion of relevant content within a given topic area. We intentionally limit our review to the specific context in which tests are used to make decisions about whether a practitioner has demonstrated competence to practice in a given profession or to use a specific practice-related designation. As such, the content of the chapter is presented with the intent of meeting the following goals:

- providing a clear presentation of the issues that historically have been and currently are most relevant for defining use(s) for developing, scoring, evaluating, and defending scores from licensing and certification tests;
- describing the legal precedents that have continued to shape standards for licensing and certification tests;
- presenting a detailed explanation of the foundational measurement issues that guide development and use of tests for licensing and certification (e.g., considerations of fairness, reliability, validity, establishing performance standards); and
- discussing related factors including policy considerations, stakeholder perceptions, and antitest sentiments.

We conclude our introduction with an overview of the major sections of the chapter: "Rationale and Historical Perspective," "Fairness and Legal Issues," "Test Development," "Reliability," "Validity," "Standard Setting," and "The Future of Assessment for Licensing and Certification."

1. “Rationale and Historical Perspective”: In this section, we begin by describing the historical context for licensing and certification assessments and explore how perspectives on these types of assessments have developed and changed over time. Content represented in this historical background includes a discussion of different types of tests and their uses, with a focus on licensing and certification, an examination of how licensing and certification practices differ in different contexts, and a focus on how assessment for licensing and certification has changed since the publication of the prior *Educational Measurement* volume in 2006.
2. “Fairness and Legal Issues”: This section reviews important considerations with respect to fairness and legal issues in credentialing contexts. The primary focus is on the legal bases that would allow for challenging testing practices and/or outcomes in both licensing and certification contexts (e.g., Title VI and VII of the Civil Rights Act of 1964; Equal Employment Opportunity Commission, 1985). We also consider fairness from the perspectives of mitigating risk for bias during test development, testing accommodation policies, and the implications of the outcomes of accommodated and nonaccommodated test scores.
3. “Test Development”: This section describes test development considerations that are particularly relevant to licensing and certification assessments. These include a discussion of the importance of clearly specifying the purpose of the test and focusing on test development as a data-driven process informed by an understanding of the knowledge and skills that one must demonstrate in the context relevant to the specific credential. Recent research on advanced test development approaches such as automated item generation and automated test assembly are included, as is a discussion of innovative item types and issues surrounding the provision of performance feedback.
4. “Reliability”: This section addresses relevant issues that are unique to or require special emphasis in licensure and certification contexts. Content includes discussion of several key topics, including classification consistency and accuracy, with a particular focus on false-positive and false-negative errors and the impact of these factors on policy decisions, as well as a review of measures of classification accuracy in the contexts of both traditional and cognitive diagnostic assessments. We also discuss the special issues in classification accuracy associated with allowing failing test takers to retake an examination.
5. “Validity”: In this section, we describe validity issues that are unique to or need special emphasis within credentialing contexts. We highlight Kane’s (2006) validity framework to discuss the importance of a multifaceted and thorough approach to collecting validity evidence in professional testing environments, discuss the particular importance of the extrapolation and interpretation components with respect to evidence for performance in practice, reflect on the theoretical underpinnings of why assessment may be important in the professions, and discuss threats to validity such as security-related issues.

6. “Standard Setting”: This section addresses issues that are central to standard setting in assessments used for licensing and certification, including a brief discussion of the content-based methods typically used in credentialing contexts, considerations related to the variety of activities that comprise the overall standard-setting process, and a discussion of the role of judgment in standard setting.
7. “The Future of Assessment for Licensing and Certification”: In the final section, we discuss what is to come in the domain of assessment for licensing and certification. This includes a review of major milestones that have influenced the practice of assessment for licensing and certification, such as the availability of new technology, reconceptualization of how competence is developed and demonstrated, and the rise of antitesting and antiregulation sentiments.

RATIONALE AND HISTORICAL PERSPECTIVE

Licensing and certification are based on a foundation that continues to evolve as professions seek to distinguish themselves through designation processes for both the profession and individuals. This history goes back hundreds of years. Some of the earliest records of licensure requirements are presented by Schmitt (1995), who described when tariffs were imposed on medical practitioners as part of an effort to regulate practice; these early requirements date to approximately 2000 B.C.E. (and are analogous to the requirement that exists in many jurisdictions of registering and paying a fee for a business license). Some of these fees were levied by governmental agencies, but as noted in the following paragraphs, professions also had an interest in controlling access to the profession itself. During this same period, the Chinese civil service began periodic testing to evaluate fitness for remaining in designated positions (see Dubois, 1970, and Clauser et al., this volume). Evidence of these efforts to regulate or monitor professions were related to healthcare, which contributed to the concern for protection of the health and welfare of the public. However, motivations for developing these systems were neither entirely altruistic nor fully driven by governmental or regulatory bodies.

The guild systems that emerged in Europe between the 1200s and the 1400s established rules for membership, expectations for education or training through apprenticeships, and member-determined expectations for entry (Schmitt, 1995). In this regard, these systems may appear analogous to contemporary professional associations that sponsor programs, but the original purpose of these systems was more about control of the profession and commerce than about ensuring minimum competency of individuals entering the profession. From an economic perspective, this control over the profession also would have influenced prices and wages. This initial shift from government-driven oversight to profession-driven monitoring represents an early phase in the movement toward more professional self-regulation. As with other policies that have used testing as part of oversight or accountability, licensing and certification programs have experienced periods of greater or lesser support for their efforts.

One reason for the reduction of oversight of certain professional licensing requirements in the mid-1800s was that there was a perception in some fields—such as medicine and law—that the schools training these practitioners were of a higher quality than had been observed at the outset of the licensing legislation. Not surprisingly, the growing recognition of supply and demand principles led to the opening of more schools to meet the increasing demand for training. However, by the late 1800s, the expectation that candidates from all training programs were equivalent had waned, as did the quality of practitioners (due to the increase in the population; Schmitt, 1995). Since then, the number of professions regulated by states or monitored by professional associations has continued to grow as professions and roles have continued to evolve.

Discussions among professional training or educational programs, credentialing bodies, practitioners, and the public have continued for decades, along with considerations for why some licenses or certifications exist. Administrators and faculty at many professional training programs often take the position that graduation from an accredited training program should be sufficient for candidates to claim that they have achieved minimum competence in the knowledge, skills, and abilities required for the credential. There have been criticisms of this perspective, however, including the fact that training programs may have incentives to retain students that are not based on competence. Instead, they are perceived to be based on factors such as the need to maintain tuition and related revenue, along with program accreditation requirements that might rely on test information to inform outcomes assessment. Concurrently, there are practitioners and policy makers (e.g., legislators, licensing board members, certification board members) who would claim that these conflicts of interest are precisely why an independent demonstration of competence is needed. Criticisms of the practitioner stakeholders often are consistent with those levied against the historic guild system, where there may be perceptions that the licensure or certification process is used as a barrier to entry of a profession and that it limits opportunities for otherwise qualified candidates, particularly those from underserved communities. Because the influence of licensing and certification testing impacts millions if not billions of people, concerns related to these perspectives are important considerations for the measurement community.

Focusing on licensure to illustrate the point, Shimberg (1982) noted that more than 800 professions require licenses; significant expansion in licensing requirements began in the early 1900s with the proliferation of licensing agencies established through jurisdictional legislation. Morath (2015) subsequently estimated that the number of licensed professions had increased to more than 1,100. When applied to the United States workforce, this represents approximately 30% of workers (Kearney et al., 2015). The increase is driven partly by the creation of new professions and roles that did not previously exist. In addition, professions have sought to increase credibility and add a layer of protection through legislative intervention. The public protection goal of professional licensing and certification serves as a powerful reference point for professions seeking both visibility and a modicum of protection from practitioners who do not meet a standard of minimum competence for professional practice.

Considering the now thousands of professions and specializations that are licensed or certified, the meaning of some of these credentials is at risk of being devalued. As an example, the plaintiff in a legal case in Utah challenged the Barber, Cosmetology/Barbering, Esthetics, Electrology and Nail Technology Licensing Board for the right to provide African hair braiding services without a license after being ordered to cease and desist from offering the service (Romboy, 2012). The court in this case noted that the specific hair braiding service was beyond the scope of the public protection interests of the licensing board and supported the plaintiff's right to be able to practice without a license. Similarly, there are instances where licensing and certification programs have been eliminated. For example, in 2019, Texas removed licensing requirements and related laws for plumbers (Ura, 2019). This decision was related to an effort to reduce barriers to entry into a trade, but unintended consequences for the public may result from making such a decision without consideration of the need for additional independent verification of the competence of individuals offering these services. These examples illustrate instances where licensing boards have potentially exceeded the scope of their public protection charge or where support for a licensing requirement has eroded (affirmatively ending a program) or become inactive (allowing a program to sunset through lack of renewal or reauthorization).

Licensure and Certification Processes

Licensing and certification are the two most prevalent examples of formal recognition associated with an occupation, role, or profession. Although we discuss them collectively, the requirements, interpretation, and use for these and other professional credentials are unique to the defined eligibility, demonstration of competence, and expectations for maintaining the credential. A first step in designing and developing a testing program involves beginning with a clear statement of purpose. As will be discussed, terminology associated with licensing and certification testing can quickly expand beyond the historical definitions. For the sake of consistency in this chapter, we will use licensing and certification terminology, but additional types of professional credentials will be discussed when relevant, along with how they connect to these more common forms.

Just as we recognize a range of achievement levels in education and a variety of roles within employment settings, it stands to reason that licensure and certification would reflect a similar diversity. This variation may be seen not only in the reasons for establishing specific licenses or certifications, but also in how scores are used and interpreted in relation to those purposes. For credentialing programs, the primary claim relates to public protection: Candidates who receive the credential meet the minimum standard for entry to the profession or ability to use the designation(s) afforded by the credential. This claim can be distinguished from a claim for uses of certification credentials in a work setting that might be used for employment selection or retention. The concept of public protection stems from the desire—of judiciary, legislative, or regulatory authorities, as well as professional associations—to help members of the public distinguish

between qualified and unqualified practitioners. For licensed professions that are directly associated with potential risk to the public (e.g., architecture, aviation, health-care, law), actions of an incompetent practitioner can produce lasting harm.

In a more historical context, the concept of public protection often is closely associated with professional licensing and certification. However, particularly in certification, the purposes have expanded beyond this core purpose. For example, some certification programs were created to recognize individuals who have demonstrated specialized knowledge, skills, and abilities in a particular domain or subdomain within a profession (e.g., specialty certification, subspecialty certification). Other certifications may function more like educational assessments and are aligned with curriculum, instruction, training, or experience and designed to serve as evidence of knowledge or skills that would be needed or desired by employers (e.g., assessment-based certificates). Even for industry certifications that are based on experience or lack formal education eligibility requirements, the underlying emphasis remains on learning and the acquisition of knowledge and skills needed to demonstrate job-related competencies. Additional certification programs have expanded more into areas such as stackable credentials, microcredentials, and badging. Although these types of designations have become more prevalent and are increasingly visible, there is little consensus on quality standards and how they are to be designed, developed, evaluated/validated, or maintained. Therefore, though we note that this is an emerging area to continue to monitor, the breadth of licensing and certification testing continues to be guided by industry standards for quality.

Distinctions between goals of public perception and market recognition are important to remember when thinking about the value of these processes for stakeholders. From a candidate's perspective, the purpose of seeking a license or certification may be to demonstrate sufficient competence to be eligible to practice in or use a specific designation within a profession. Beyond minimal requirements for a role or profession, there often are market-based incentives for candidates to demonstrate deeper or more specialized competence that help them to distinguish themselves from colleagues. In addition, certain credentials may be used as part of employment eligibility requirements; this has implications for necessary validity evidence as well as legal expectations.

Link to Employment Testing

Because results from credentialing examinations may be required as part of employment eligibility processes, the question of whether these examinations could potentially be interpreted as employment tests that support selection purposes becomes relevant. Phillips (2017) discussed these implications based on the *Gulino v. Board of Education* (2002) case in New York. In that case, the court viewed the licensing requirement for teachers as a *de facto* employment test because teachers who had been employed without certification were either not promoted, not hired for full-time positions, or terminated based on their certification test performance. Although the primary

purpose of a credentialing examination may be to differentiate between candidates who are minimally competent and those who are not, integration with the employment selection process indicates a clear—though perhaps not universal—secondary use. The *Standards for Educational and Psychological Testing* (Standards; American Educational Research Association [AERA] et al., 2014) distinguish between tests used for selection, placement, and promotion in employment contexts and those used for licensing and certification, which are designed to determine whether individuals meet established criteria for professional practice. Selection tests within employment settings may be used to rank order or categorize candidates in the hiring process. This is particularly useful if the number of available positions is limited. To accomplish this goal, the test needs to yield scores that support those types of decisions. This distinction between rank ordering candidates for selection, promotion, or retention purposes versus a broader claim of minimally qualified or not has implications for test design, development, and validation, as well as the evidence needed to support legal defensibility of the resulting scores (see Ercikan & Solano-Flores, this volume).

Because licensing and certification programs often have educational components as part of their eligibility criteria, the intersection of expectations for testing in these areas along with employment considerations requires clear distinctions among purposes and heightens the importance of transparency with respect to associated limitations. Confusion among stakeholders increases when programs' purposes are not clearly delineated. The expansion of purposes can raise challenges for licensing and certification programs in terms of both validation evidence and defensibility. For example, within student achievement testing in the United States, we have observed policies that incorporate results for purposes of both school and educator accountability. In higher education, results from licensing and certification exams often are used as an indicator of outcomes for purposes of program accreditation. Within licensing, provisional or conditional licenses may be granted in emergency situations that then require successful examination performance as a condition of continued employment. Often seen as intuitive by policy makers or sometimes even by programs seeking to provide evidence for or expand credibility of its credential, additional interpretations or use cases should be critically evaluated prior to implementation.

Program Sponsors

Licensure and certification programs have a wide range of sponsors. Agencies, associations, or organizations that grant the credential may assume responsibility for development and validation of the associated tests (e.g., American Board of Internal Medicine, National Commission on Certification of Physician Assistants). At the same time, some professions have formed associations or collaboratives where one or more intermediary agencies assume responsibility for development and validation of the tests (e.g., National Council of Architectural Registration Boards [NCARB], National Board of Medical Examiners, American Board of Dental Examiners, Federation of State Boards of Physical Therapy). In these instances, test users (e.g., state licensing boards,

regulatory agencies) may contribute to the process and then agree to accept or recognize the results of the tests to support consistency and portability.

For some credentialing contexts, membership associations may serve as the sponsor for the program (e.g., American Physical Therapy Association for physical therapy specialties, National Strength and Conditioning Association for strength and personal training). In some professions, this may be a single professional association; in others, there may be multiple membership associations or programs that compete for credibility and market share among members of the eligible population. In addition, some credentialing programs are sponsored by organizations that then seek to certify or confirm qualifications of individuals to perform specific roles within divisions of the parent organization (e.g., Armed Services, Federal Bureau of Investigation, National Security Agency within the U.S. Department of Defense). However, these credentials sometimes begin to have value in the workforce or labor market beyond just the requirements of the sponsoring organization. For example, industry certifications developed by companies in diverse economic sectors such as finance, healthcare, or technology are valued by a range of employers, internally and externally, that use products and services from these companies. Within employment settings, companies and governmental agencies may develop internal credentials that are used to support nonselective promotion or wage decisions.

The distinctions among government-regulated, profession-driven consortia or combinations of these processes are discussed in the next section. As noted previously, the primary differentiator is whether holding the license or certification is mandatory or voluntary for practice, with these differences creating more confusion among stakeholder groups.

Mandatory Professional Credentials

A license likely is the most frequently cited instance of a mandatory professional credential. Well-known professions such as architecture, accounting, law, medicine, nursing, and clinical psychology require practitioners to obtain a license to practice or represent themselves in their profession. What may be less well known, however, is that many jurisdictions also require professionals such as auctioneers, barbers, contractors, cosmetologists, real estate agents, and security guards to obtain a license to practice.

Licensure generally is a jurisdictional legislative responsibility. In the United States, there are some exceptions, including the Federal Aviation Administration's oversight of private pilot licenses and the judiciary's responsibility for admission to the bar for lawyers. As such, there may be jurisdiction-specific requirements for eligibility, demonstration of competence, and maintenance of the license. As described earlier, the focus of licensure is public protection and is based on the rationale that a governmental agency can serve as a dispassionate intermediary between the candidate and the public in evaluating minimum competence. However, even these agencies will necessarily rely on or seek input from members of the profession to help define expectations for entry-level practice.

Although having a license may be necessary to practice in a given field, it does not ensure employment or even employability because of other factors that may go into selection. Further, there are exceptions even to the *mandatory* requirement. For example, some emergency situations (e.g., severe shortages of practitioners, pandemics, natural disasters) may necessitate a temporary relaxation of requirements to grant a provisional or conditional license so that more practitioners can be mobilized. A credentialing examination may serve as an eligibility criterion within the hiring process, similar to the expectation that a valid driver's license is a prerequisite for applicants for a delivery driver position. However, scores from these examinations are not designed to predict successful performance on the job or distinguish among applicants who meet minimum eligibility requirements. In contrast to claims made about scores for employment selection tests, validity evidence for a licensing or certification exam should be sufficient to support the inference that a candidate has met the standard for minimum competence. In addition to this initial demonstration, some credential holders are expected to maintain the credential in good standing; this may involve activities such as continuing education, retesting, or contributions to the profession. The purpose of this practice is to ensure that licensees have a level of continued competence to maintain currency in the profession. This maintenance component often is what distinguishes professional credentials from certificates or industry certifications of products or services that are intended for specific skill sets (e.g., technology certifications, trade certifications, course-based certificates). However, it is important to note that some of these industry certifications do have a maintenance or continuing education requirement.

Voluntary Professional Credentials

In instances where someone can practice in a field without a license, a certification can serve as a point of distinction for the public. Because a voluntary certification may be sponsored by an association, a private organization, or the profession itself, there are fewer barriers to creating a certification; this can be offset by the challenge of building credibility for the value of the credential. Licensed physicians can work as general practitioners, but it is common to seek one of the many specialty/subspecialty certifications offered by American Board of Medical Specialties (ABMS) member boards (e.g., Emergency Medicine, Family Medicine, Internal Medicine, Obstetrics/Gynecology, Pediatrics). Within some jurisdictions, lawyers may seek specialty certifications in criminal law, estates and trusts, family law, or intellectual property. Individuals in the financial services sector may strive to obtain recognition through programs like the Chartered Financial Analyst Institute or as a Certified Financial Planner. These examples illustrate the range of options available to individuals seeking opportunities to voluntarily demonstrate competence within a specific professional domain or subdomain.

In contrast to the jurisdictional oversight of licensure programs, certification programs could be developed by a professional association or a company that offers a product or service on which someone desires to be certified. As a result, there is a wider

range of program designs and methods for implementing them in practice. There may be some level of competition within the certification program space where multiple providers offer credentials (e.g., crane operators, personal trainers, food protection). In these instances, it can be more difficult to ensure independence and mitigate conflicts of interests in governance policies with respect to program development and implementation.

In terms of test development and validation, certification programs are like licensing programs in that the primary claim relates to competence for practice. In addition to eligibility or prerequisite requirements, the demonstration of competence is based on the knowledge, skills, and abilities related to the type of practice implied by the certification. In industry, this may be defined in terms of a level of mastery; in other professions, it may relate to competence in some area of specialization. Earning a voluntary certification does not ensure employment but may be used in the employment process as part of preferred qualifications of applicants. In contrast, if a certification is used as an employment eligibility requirement, it could be interpreted as shifting the certification from voluntary to mandatory.

Although not universal, many professional certification programs have maintenance of competence requirements like those seen in licensure. These requirements will be unique to the program but may include such activities as continuing education, contributions to the profession, or retesting. In a separate section of this chapter, we further explore how some healthcare certification programs offer maintenance of competency assessments that blend learning and assessment. In addition, some certification programs have depended on assessment to inform continuing education recommendations for certified individuals. Related to licensure, there is some debate about whether maintenance of competence requirements should include a demonstration of *continued* or *enhanced* competence. From a measurement perspective, the question is related to the intended interpretation and use of the evidence from the maintenance of licensure process.

Program Stakeholders

Stakeholders for licensing and certification programs are varied, and each has its own perspectives and interests associated with the credentialing process. The public often is noted as the primary stakeholder because public protection tends to be the principal reason for the existence of these types of programs. Clearly, it is important to protect the public from an incompetent physician, teacher, attorney, engineer, or crane operator. In general, when there is a jurisdictional license involved, there has been a decision that incompetent practice could harm the public and that market forces may not offer sufficient protection. This may be particularly true for more technical or specialized domains where members of the public would not be able to differentiate competent from incompetent performance. Beyond initial licensure, another public interest may be served by recognizing individuals who have demonstrated more specialized competence within their field to distinguish themselves from others who have not achieved an additional credential.

Another key stakeholder group comprises the candidates or applicants seeking the credential. These individuals have a personal interest in achieving certification and a legitimate expectation of fair access to professional practice. Their perspective may shift after becoming certified, as the same standards that once posed a barrier now serve to uphold the value and credibility of the credential they hold. In many certification programs, candidates also are connected to professional associations or organizations that define standards for the profession. These groups represent another type of stakeholder with a dual responsibility: to protect the public interest and to serve their members. At the same time, they may have an organizational self-interest in maintaining the perceived value of the credential and sustaining membership, which supports their continued existence.

The training programs that prepare candidates for practice comprise yet another important stakeholder group. The nature of these programs varies across professions. In the case of medicine, the stakeholders in this group are medical schools. For certification, the stakeholders are residency programs. In both cases, access to the credential is important both because individuals are attracted to the program based on their belief that they will succeed in receiving the credential and because accreditation of the training programs may depend on satisfactory performance using this same metric.

Within the healthcare community, insurance providers are another stakeholder group. They may require medical professionals to have certain credentials to provide treatment and may have different pricing or reimbursement tiers for different types of credentials. Possession of the credential also may serve as a form of risk management by having third-party agencies or organizations recognize someone as having a particular skill set. In this instance, it distributes the risk across the insurance company, the provider, and the credential-granting entity.

For employers, credentialing adds a layer of quality control and external attestation of competence that may not be able to be discerned on one's own. Employers in the public and private sectors are stakeholders of credentialing programs and sometimes sponsors of these programs. For example, employers may require a given credential as an expectation for consideration of employment. In these instances, employers are placing trust in the credentialing body and the credential itself as being able to support associated claims about competence. Similarly, labor unions may want to use credentials not only for helping to establish quality standards for their members but also as part of a negotiating strategy to support increased leverage in advocating for the competence of their members.

Additional Evolving Practices

Whereas the fourth edition of *Educational Measurement* highlighted some of the measurement approaches in computerized scoring of performance assessments (e.g., NCARB), challenges associated with staying current with respect to changes in the technologies used in practice have altered the strategy for some of these programs.

For example, NCARB removed performance items (i.e., vignettes) that required candidates to use a generalized but simplified design software to respond to tasks. This decision was driven both by changes in practice regarding the roles responsible for these activities and by the job relatedness of the software that was being used. In response to those concerns, they opted for a more flexible strategy that permitted continued measurement of the important content and cognitive processes associated with being an entry-level architect and provided better alignment of the measurement with the task (see NCARB, 2023a; NCARB, 2023b). This approach to designing an examination—which is less dependent on a specific technology and more responsive to the knowledge and skills required by the profession—is becoming more prevalent in practice.

Similarly, as noted in the fourth edition, a content-neutral demonstration of professional skills has been part of many jurisdiction bar examinations for years. In this approach, candidates (or applicants) are provided with a library of resources and are expected to complete a task that requires them to engage with these materials and respond in a job-relevant manner. The emphasis on skills has taken on new meaning as the number of practice areas has continued to expand. Based on national job analysis research (National Conference of Bar Examiners, 2020), more than 25 different practice areas were noted by survey respondents, suggesting that the scope of practice is diverse. An additional unique feature of the bar examination is that, rather than being overseen by the legislative branch of government as is typical for other professions, it is the only licensure process overseen by the judicial branch of government (within jurisdictions).

As we hope has been demonstrated, there is a long and complex history behind the field of assessment for licensing and certification. Many advances have been made with respect to understanding the complexities of credentialing contexts and how to advance both theory and practice, yet many challenges still exist. The next section reviews topics relating to fairness and other legal considerations that can provide more insight into some of the issues and challenges that have faced this field in recent years.

FAIRNESS AND LEGAL ISSUES

The high-stakes nature of decisions made based on scores from credentialing examinations places additional responsibilities on programs that develop and maintain these examinations. Specifically, in addition to the professional expectations articulated in the *Standards* (AERA et al., 2014), programs and practitioners should be aware of the legal frameworks and case law that may interact with theory and practices within the measurement community. Phillips (2017) organized her comprehensive discussion of legal challenges that have occurred in licensure and certification using the following framework:

1. *Protecting the public*: Conditions under which it is fair to require candidates to demonstrate minimum competence for defined knowledge, skills, and abilities.
2. *Testing accommodations policies*: Defining criteria for disabilities and reasonable accommodations and ensuring that the meaning of the license or certification is not changed through modification of the construct.
3. *Test security policies*: Procedures for maintaining confidentiality of test content, ensuring valid score interpretation through detection and enforcement of test security violations.
4. *Test construction procedures*: Options for licensing and certification programs pressured to adopt item selection criteria that minimize differential performance between majority and minority candidates.

We discuss some of the recent case law associated with the first three of these areas and refer interested readers to Phillips (2017) for discussion of some of the historical background associated with the fourth area as it relates to legal precedent that conflicts with what often are considered professional expectations and best practices within measurement. Though this section largely focuses on case law covering issues related to fairness, the topic of fairness also is explicitly discussed relevant to test accommodation and score interpretation and is included as central to test development, reliability, and validity considerations. For a broader discussion of issues related to fairness, the interested reader is referred to Zwick, and Rodriguez and Thurlow, both this volume.

With a focus on U.S. jurisdictions, federal law forms the basis for most of the challenges described in this section. Most legal challenges to licensing and certification programs point to alleged violations of the 14th Amendment to the U.S. Constitution, Title VII of the Civil Rights Act (1964), Section 504 of the Rehabilitation Act (1973), the Copyright Act (1976), and the Americans With Disabilities Act (1990). The courts have relied on the *Uniform Guidelines on Employment Selection Procedures* (Equal Employment Opportunity Commission, 1985) for interpreting expectations for licensing and certification litigation. These legal expectations generally are intended to guard against discrimination of protected classes that include the demographic characteristics of age, disability, ethnicity, race, or sex. It is important to note that these legal expectations are not universal and some of the case law discussed in this section may not warrant action or be part of legislation in other countries. As such, readers should consider the jurisdiction-specific expectations applied to their licensing, certification, or equivalent program (e.g., qualification, accreditation, registration).

As noted earlier, there are documented examples of licensing boards overreaching in terms of their charge to protect the public (e.g., the plaintiff offering African hair braiding services). Similarly, decisions like the one against the North Carolina Board of Dental Examiners (2015) indicated that the courts recognize that there is a limit to how a licensing board defines its need for oversight and public protection. This case began with a complaint filed with the Federal Trade Commission by owners of teeth-whitening kiosks after the state board of dental examiners claimed exclusive domain under their practice

act and told them to cease and desist. The Federal Trade Commission concluded that the interpretation was an overreach of the practice act and ruled against the board, citing that, even within a dentist's office, nonlicensed staff members administered and monitored teeth-whitening procedures, thereby suggesting that the risk to the public was limited. Within Phillips's (2017) organizing framework, this illustration would fall within challenges related to assertion of public protection. Challenges to licensing and certification examinations may not come through judicial processes such as lawsuits because legislative (e.g., House, Assembly, Senate) and policy bodies (e.g., licensing or certification boards) seeking to sunset (Drew, 2016) or remove (Taylor, 2017) requirements may be as effective as litigation.

Because the treatment of the historical foundation for legal challenges for licensing and certification programs is articulated in greater detail by Phillips (2017), B. E. Clauser et al. (2006), Phillips and Camara (2006), and Ercikan and Solano-Flores, and Zwick (both this volume), we have limited our discussion to key decisions that have occurred within each of these areas since the publication of the fourth edition of *Educational Measurement*. This is not a comprehensive review of all decisions but rather illustrates the types of challenges that have occurred in licensing and certification.

Protecting the Public

Because of the potentially broader implications for licensing and certification testing programs, the *Gulino v. Board of Education of the City School District of the City of New York* (2002) case is of particular interest for the measurement community. In the Gulino case, the court interpreted a teacher certification examination as having properties of an employment test because of how the state permitted provisional or temporary licenses that allowed teachers to teach. This temporary license included an expectation for completion of licensure examination requirements within a specific time frame. Because passing the examination became a condition of continued employment in a given role, the court viewed this as a de facto employment test. When examinations are mandatory for employment eligibility, as is the case for licensing and many certification examinations, the Gulino ruling could be extended to the broader family of licensing and certification examination programs in which passing the examination is a condition of initial or continuing employment. Litigation on the case has been ongoing for more than 2 decades.

The original action was filed in 1996, when a group of African American and Hispanic teacher candidates alleged that the teacher certification test discriminated against racial minority candidates because passing rates for both racial/ethnic groups were below the 80% rule that defined disparate impact under the *Uniform Guidelines*. The 80% rule means that a protected class has an observed passing rate that is less than 80% of the passing rate of the majority group. The lawsuit was certified as a class action in 2001. As it relates to test development and validation, the court's interpretation of necessary evidence to support the test's continued use was important. The court ruled that five sources of evidence were needed to support the job relatedness of an employment test.

Specifically, these sources of evidence were (a) a suitable job analysis, (b) competent test development, (c) test content related to job content, (d) tested content representative of the job, and (e) an appropriate performance standard (i.e., passing score) for selecting competent applicants. The court held that the new test was not properly validated because the vendor did not conduct a rigorous job analysis, did not competently develop the test, tested content unrelated to and unrepresentative of the job of teaching, and applied a passing standard unrelated to teaching competence (Gulino, 2002).

More specifically, the court found that the vendor's test framework—developed by reviewing teacher education materials, consulting with education experts, and surveying teachers and teacher education faculty—was flawed because the vendor: (a) did not start with a task analysis of teaching; (b) had not documented the reference materials or persons interviewed; (c) had pilot tested items on college students rather than on working teachers; (d) had failed to link the tested content to minimum and representative content required for teaching competence; and (e) set the passing standard based on a small subset of items with no definition of minimum competence or data relating test scores to student outcomes. Thus, the court concluded that the board violated Title VII when it required teachers to pass the new test to obtain permanent licenses. The decision was affirmed on appeal.

Several lessons emerged from this case. An important lesson was that even though the *Standards* (AERA et al., 2014) is intended to serve as a primary guide for the testing profession, the court in this case did not recognize it as such. Further, because the court appointed a neutral expert who prioritized the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) along with the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1985), the value of the *Standards* in this proceeding was further diluted. Although not intended to be prescriptive, if guidance from the *Standards* does not align with legal expectations, practitioners will be exposed to risk. Although the *Standards* includes a caveat about its utility in legal proceedings that may supersede the profession's expectations, these legal standards often are not widely known, nor are they necessarily consistent. Different jurisdictions may have different expectations, so beyond some core expectations that would be articulated in federal law (e.g., U.S. Constitution) or regulations (e.g., Civil Rights Act of 1964 as amended in 1991), the legal standard may not become fully evident until it is settled by the Supreme Court. The court's rationale for interpreting the state's credentialing exam as an employment test was that it was used as a condition for continued employment following issuance and potential extension of a temporary or provisional license, not just eligibility for employment. Therefore, without passing the examination, an applicant would not have been eligible for a permanent license. The focus of the proceedings then centered on the job analysis evidence and whether there was a clear connection between knowledge, skills, and abilities that an educator would need to know in practice.

Although evidence from the state and most evidence collected by the vendor from practicing educators and faculty who train educators suggested that there was a connection, the court's expert disagreed and offered an opinion that suggested a specific methodological preference for what would constitute an appropriate job analysis study. The perspective offered by the court's expert focused on the incremental validity evidence provided by the test in question and whether all educators should be required to demonstrate this level of minimum competence. During the hearing, the plaintiff's attorneys and the judge questioned whether the range of content represented on the test was job related and important for all educators. The intuitive perspective of the judge was that content areas such as fine arts would not apply to what a science teacher would need to know and be able to do (*Gulino, 2015a*).

Concurrently, the state had been in the process of redesigning and developing the examination with a focus on making changes to the content specifications. The revised examination recharacterized the construct as inclusive of literacy and writing abilities rather than the range of liberal arts skills that were defined in the prior version. This new exam also was challenged as part of the proceedings in the case as having adverse impact and not being properly validated. In this instance, the state's test vendor contracted with an external organization to conduct a practice analysis. The extensive study evaluated the knowledge, skills, and abilities of educators to determine whether the content of the examination was consistent with job-related expectations.

Again, the court's appointed expert concluded that the examination was not properly validated for reasons almost identical to those advanced for the prior iteration of the examination in the broader program. However, in this instance, the court's ruling for *Gulino* (2015b) yielded a different outcome in that the court accepted the content evidence for the program. Although the methodology for evaluating content evidence for these examinations was different, they were both within a common set of validation activities conducted for credentialing examinations. It is possible that the court's acceptance of the evidence of examination content for the new examination was based on an intuitive understanding of content that educators would need to know and be able to use. In other words, it was more difficult to argue that the literacy skills of reading and writing content areas were not important for all educators in practice, as opposed to the diversity of content represented in the prior version. However, the opinion of the court's expert likely leaves room for additional litigation on this issue because the implications for the testing program and its users are substantial. The ongoing litigation costs and questions regarding how to interpret the court's rulings in this case have continued to have ripple effects for other educator certification programs as well as for the licensing and certification community more broadly.

Testing Accommodations Policies

The Americans With Disabilities Act (ADA, 1990) requires that licensing and certification programs make decisions about accommodations for candidates with disabilities (see Rodriguez & Thurlow, this volume, for further discussion of fairness

issues related to test takers with disabilities). To support a violation of the ADA, plaintiffs need to provide evidence that there is a qualifying disability and that a reasonable accommodation was denied (*Cox v. Ala. State Bar*, 2004). The three elements of a qualified disability are (a) an impairment (b) that substantially limits (c) a major life activity. As it directly relates to some of the knowledge and skills measured in licensing and certification tests, it includes many prerequisite abilities such as literacy and communication skills that generalize across professional domains.

Accommodation, Not Modification

It is important to differentiate between what is intended with an accommodation and something that would modify or change the construct that is being measured. In the instance that providing an accommodation would fundamentally change how the job-related knowledge or skills are demonstrated, a candidate does not need to be permitted to receive the accommodation under the ADA. It then becomes important for a licensing or certification program to define the critical knowledge, skills, and abilities that are part of the intended construct. Phillips (2017) suggested that licensing and certification programs that use simulation or more real-world measurement approaches may not be able to offer accommodations as easily as examination programs that focus more on knowledge or application of information without potentially modifying the intended construct.

Although accommodations often are interpreted within the measurement community as facilitating opportunities for candidates with disabilities to demonstrate their skills, there is an important distinction in stakeholder groups' perceptions. Phillips (2012) noted that advocates for disabled candidates may interpret increased access as applying to opportunities within a profession or occupation; this is notably different from attempting to facilitate greater access to the examination. She further makes the point that if the accommodation is construct relevant, the interpretation of the candidate's job-related abilities may be called into question as a result of the shift in the meaning of the score and resulting decision.

The interesting court ruling in *Palmer College of Chiropractic v. Davenport Civil Rights Commission* (2014) raises the question of how job relatedness is interpreted, and it extended the question about what types of accommodations are reasonable. In this case, the court ruled that the college was required to provide a visually impaired student with a sighted assistant for reading and interpreting radiographs (i.e., X-rays). The question of the job relatedness of being able to read, interpret, and diagnose radiographs for the purpose of developing and implementing a treatment plan is an important one. The court's ruling suggests that it is a skill that in practice can be delegated to another role in support of the chiropractor role. If this is the case, then it is a reasonable interpretation. If not, however, then the court in this instance has appeared to define expectations that may run counter to the legislative practice act that defines the scope of practice for the profession in the jurisdiction.

Evaluating Reasonableness

Part of the ADA's expectations for accommodations rests in the interpretation of what constitutes a *reasonable* accommodation, recognizing that there are some limits to the interpretation. In *Kelly v. W. Va. Bd. of Law Exam'rs* (2010), the plaintiff (Kelly) was diagnosed with a reading disability after being accepted to law school and received a time-and-a-half accommodation for exams. He then transferred to a new law school, where he received a double-time accommodation. He did not receive any accommodations during college or for his admission exams.

For the West Virginia bar examination, Kelly requested double time but was granted a time-and-a-half accommodation. Although it was not part of the bar exam, Kelly also passed the required professional responsibility exam within the licensure process for lawyers without an extended time accommodation. A member of the licensing board testified that it was a job-related skill for a lawyer to be able to work under time constraints and that most applicants used the full testing time, with not everyone finishing. The plaintiff failed the West Virginia exam twice with the time-and-a-half accommodation and then successfully appealed to Kentucky's Board of Bar Examiners to provide a double-time accommodation for its bar exam.

In evaluating the differing opinions of expert witnesses for the plaintiff and the board, the court held that the board's position was more credible given the plaintiff's prior education experience and related test score performance that were provided as evidence of the need for the accommodation. The court further held that providing more time than was necessary would shift the impact of the decision from being a disability-based accommodation to being an advantage over other applicants. As part of its deliberation, the court considered the historical pattern of accommodations for the plaintiff: specifically, that there were no accommodations provided until law school. As a result, the court interpreted that the request was unreasonable and not required under the ADA.

Currier v. NBME (2012) was another case that focused on the definition of a reasonable accommodation. The context for this case was Step 2 of the United States Medical Licensing Examination (USMLE); the two-part accommodation request was for (a) extended time for diagnosed learning disabilities and (b) additional break time for the candidate, a nursing mother, to pump breast milk. In response, the National Board of Medical Examiners (NBME) offered to provide a double-time accommodation for the learning disabilities (dyslexia, attention deficit hyperactivity disorder), administering the exam over 2 days, in a separate room. To respond to the candidate's request for the 60 extra minutes each day to pump breast milk, the NBME offered to provide its standard break time of 45 minutes each day, along with a power outlet and the ability to bring food into the separate room that was previously offered for the double-time accommodation.

The plaintiff sued to get the extra break time to be able to pump. NBME's position was that the break times needed to be uniform for all candidates, and the combination of accommodations—double time and extra break time—made

the administration logistics untenable. The trial court sided with NBME, but on appeal, the court then sided with the plaintiff and ruled that the extra-time accommodation should be provided. With that additional court-ordered accommodation, the plaintiff failed the examination; she then passed the examination after retesting without the accommodation. As the legal proceedings continued, the court decided in favor of NBME, with the plaintiff appealing again. The appeals court ruled that nursing mothers were covered under the *state* equal rights act, but not under the ADA, based on sex discrimination. It is important to note the distinction between state and federal jurisdictions and how they may not always overlap with respect to interpretation.

Test Security Policies

Many licensing and certification testing programs have transitioned to computer-based testing. This occurs in testing centers, through in-person events, or with remote proctoring. Although early proponents of computer-based testing argued that it would increase test security because the test material could be encrypted, concerns about security of test content and validity of scores have continued to be a significant problem. Not surprisingly, this has led to litigation related to test security and administration. Here we highlight two instances of these types of cases.

In *ETS v. Hildebrant* (2007), a candidate who was taking a principal certification test had signed a candidate agreement about confidentiality, nondisclosure, and the testing conditions included in the candidate manual for the program. According to the test proctor, the plaintiff twice refused to stop writing when time was called. Following an initial warning, the proctor completed a test irregularity report to document the nonstandard condition and noncompliance of the candidate to the administration timing conditions. The candidate insisted that she had not violated the administration conditions. After an investigation, the vendor concluded that there was a violation and canceled her test score. During litigation, the court held that the candidate's unsworn, general denial was insufficient evidence to create doubt that the vendor acted in good faith. Further, the plaintiff was unable to convince the court that the proctor had a motive to lie about the violation of the administration protocol. The court held that the vendor complied with the terms of its own policies and procedures in relying on the proctor's observations and documentation of the violation. As such, the court agreed with the vendor's position that the misconduct justified score cancellation.

A second example, *NCBE v. Multistate Legal Studies* (2006), deals specifically with the security of test material. In this case, the defendant was the Preliminary Multistate Bar Review, a test preparation provider that offered review courses for the Multistate Bar Exam across multiple sites. Detection of an attempt to steal test content occurred after a proctor caught an applicant—who was an employee of the provider—attempting to remove test content on scratch paper. Following an investigation that uncovered approximately 100 stolen items, the National Conference of Bar Examiners (NCBE) brought a lawsuit against the provider for copyright infringement. In the court's

discussion of their decision in support of a ruling for NCBE, they found substantial overlap between items that were offered by the test preparation provider and the actual items from NCBE's examination.

Discussion of legal considerations that are relevant to licensing and certification tests provides some insight into areas that should be a central focus of test developers and score users because of the associated risk for legal action. In the next sections, we review three main topics that are critical in high-stakes credentialing contexts because of their direct impact on the interpretation and use of test scores: test development, reliability, and validity.

TEST DEVELOPMENT

In most, if not all, credentialing contexts, candidates seeking the credential will be required to complete at least one examination to demonstrate their competence. A passing score on this examination will serve as an indication that the test taker has the requisite knowledge and skills to competently and safely perform the duties of the profession. As described in the *Standards*, a number of specific considerations and recommendations apply to design and development of tests used for educational, psychological, or credentialing purposes (AERA et al., 2014). This section presents test development considerations that warrant particular attention in the context of licensing and certification; for a comprehensive review of test design and development more broadly, we refer the interested reader to Huff et al., this volume.

Specifying the Purpose

The general purpose of credentialing examinations is clear: to identify individuals who have the minimum knowledge and experience to perform required tasks competently and safely (Chinn & Hertz, 2010). As noted previously, the focus in licensure is on the need to demonstrate that an individual is qualified to practice in the profession; in certification, the focus may be on specific specialization within the profession or on a more advanced level of performance.

The importance of developing and publicizing a clear statement about the purpose of the overall test—or, when relevant, the main components comprising a test—cannot be overstated. Doing so serves several functions. First, it guides subsequent test development efforts by specifying what is in and out of scope for the examination. It also informs stakeholders about the nature of the examination. For example, the NCARB has publicized the following purpose statement about the Architect Registration Examination (ARE):

The ARE is designed to assess aspects of architectural practice related to health, safety, and welfare. Specifically, the ARE focuses on areas that affect the integrity, soundness, and health impact of a building, as well as an architect's responsibilities within firms, such as managing projects and coordinating the work of other professionals. (NCARB, 2023b)

Although the specifics of the examination content are not provided, the reason for administering the test and general test content are clear to test developers and other stakeholders, including candidates who will be taking the examination. According to Cizek et al. (2011), “The test development process begins and is guided by clear, explicit statements regarding the purpose of the test and the inferences that are intended to be made from the test scores” (p. 6).

One of the challenges inherent in constructing a purpose statement is developing a clear understanding of (a) what *should be* assessed with respect to the critical knowledge and skills needed by practitioners, as well as a realistic idea of (b) what *can be* assessed given the inherent constraints associated with assessment. Careful consideration should be given to the requirements of practitioners at the specific level at which the examination is targeted. This need highlights the importance of a systematic evaluation of the profession that will yield information to guide decisions about examination content (AERA et al., 2014; Chinn & Hertz, 2010; Cizek et al., 2011; Raymond, 2001, 2002) and, by extension, development of a purpose statement.

Identifying Appropriate Content

As we have mentioned, one of the central requirements for credentialing in professional disciplines is successful completion of an examination or series of examinations that is/are intended to provide insight into whether the test taker has the necessary knowledge and skills to practice in the profession safely and competently. Regardless of the discipline or the type of credentialing, the common thread among these examinations is the need to determine the content that will provide the evidence for minimally acceptable performance. According to the *Standards*, best practice for determining the content of credentialing examinations is surveying members of the profession about the knowledge, skills, and abilities that are required of practitioners to identify the actual tasks that practitioners must be able to perform competently and safely. If done properly, this *job* or *practice analysis* will collect information about what activities are required for practice and thus will inform decisions about the content that is relevant for the examination (A. L. Clauser & Raymond, 2017). It is important to note that the practice analysis does not dictate what will be included on the test, but instead provides test developers with information about the requirements of the profession; it is the responsibility of the test developer to carefully review the results to identify the content that can be assessed within the constraints of the examination format.

A license indicates that the practitioner has the knowledge, skills, and abilities required for competent and safe practice. It is not a prediction that the individual will develop those skills over time and, in most cases, it is not restricted to a set of entry-level tasks.² To use an example from outside the professions, a driver’s license allows an individual to operate a motor vehicle. It may be ill-advised for the newly licensed driver to operate a vehicle in a highly congested area, but the license does not include such a restriction. Similarly, when a state licenses someone to practice medicine, the license does not come with prescribed limitations. There are, however, circumstances where these lines are blurred.

For example, some states require passing the first two steps in the three-step USMLE to enter a residency program and practice under supervision. Similarly, residency programs typically require successful completion of those components of the examination, making those components of the test a *de facto* license for practice under supervision. In many professions, there is a presumption that practitioners will improve their skills after being licensed; advanced certification documents that improvement, but licensing tests typically are a stand-alone evaluation. It is not uncommon for language about “entry level” or “readiness to enter the profession” to be used when discussing licensure. Insofar as this characterization implies a trajectory, we believe that it may be misleading.

It is worth mentioning that practice analyses done for the purposes of occupational licensing and certification may be different than those conducted for other purposes. For credentialing purposes, the focus is on those areas that are observable and are related to “impact on public health, safety, and welfare” (Chinn & Hertz, 2010); consideration of those areas that provide insight into other professional activities, such as those that allow for personal achievement within a profession, for example, would not be relevant. This focus on the safety of the public is a thread that carries through the entire test development process.

Much has been written about practice analysis design and the relationship between the design and the ability to extrapolate data to inform test specifications (and, ultimately, examination content; Chinn & Hertz, 2010; Raymond, 2001, 2002). The following paragraphs provide a brief overview of practice analysis in credentialing as well as recommendations and cautions associated with its use (see Raymond, 2001, 2002; and Raymond & Neustel, 2006, for a more detailed review of relevant considerations).

The decision to conduct a job analysis is only the first step in the process; a series of additional decisions about the specifics of the process must follow. These decisions generally relate to the scope of the project, the overall methodology that will be used, and the rating scales that will allow for collecting the actual practice data. Subsequent decisions about use of the resulting data to inform test specifications also must be made, but careful attention to the initial design considerations will help to ensure a more seamless process of applying the data to the next phase of test development.

Numerous approaches to practice analysis are available to practitioners, and much has been written about the different methodologies and their associated strengths and weaknesses in particular contexts and for particular uses (Raymond, 2001, 2002; Raymond & Neustel, 2006). That being said, of particular focus in credentialing contexts has been the task inventory approach and, subsequently the professional practice model (Raymond, 2002).

With the task inventory approach, a list is created that outlines the tasks or activities that are performed by practitioners in a profession. The list typically would be created by a group of subject matter experts, and the results then would be included in a survey that asks a representative sample of people in the field to respond to rating scales about different aspects of the included tasks (e.g., the importance of the task for

practice or the frequency with which the task is performed). The primary benefit of this approach is efficiency: A lot of information can be collected in a short time. The breadth of information that can be collected is especially important in credentialing contexts (where readiness for performing a variety of tasks needs to be specified across a range of settings), and the data translate well to the test specification phase (Chinn & Hertz, 2010; Raymond, 2002). A disadvantage, however, is that the focus is on individual tasks rather than higher level skills such as critical thinking and problem-solving approaches. Task inventories that emphasize discrete, observable tasks may overlook the cognitive nature of many professions (LaDuca, 1994).

This approach, regarded as an overarching framework within which practice information can be organized rather than a distinct method (Chinn & Hertz, 2010), focuses on including multiple dimensions that are important to the practice. Doing so allows for creating a matrix in which the practice analysis content is included at the intersection of the two dimensions. In medicine, this might be represented as a setting-by-competency design in which the practice settings (such as inpatient wards, emergency department, ambulatory clinic) are crossed with critical competencies such as medical knowledge or communication skills. The tasks that rely on a given competency within a setting then are included in a given cell. This approach is more comprehensive than a strict task inventory and likely is more appropriate for the complex domain of professional credentialing. The framework also is consistent with the model of practice analysis proposed by Kane (1997).

The professional practice model was developed as a way to address these limitations; it focuses on the problems that practitioners will need to solve within the relevant contexts in which they occur (Chinn & Hertz, 2010; LaDuca, 1994; LaDuca et al., 1995; Poniatowski et al., 2019; Raymond, 2002).

Though a variety of acceptable options for data collection exist, research findings indicate that a more comprehensive approach may be preferable to any single identified or named method. A comprehensive practice analysis therefore might include the following elements. First, instead of looking at discrete tasks or specific knowledge, skills, or abilities, it would focus on the types of problems professionals solve and the tools and ideas they work with to solve them. Second, it would be model driven. The model could be used for specifying a preliminary domain of practice, developing questionnaires, and establishing content weights. The goal of the practice analysis would be to confirm, modify, and refine the model. Third, the information would be obtained from multiple qualified sources, depending on the type of information being sought (e.g., a large sample of practitioners for some types of information, panels of subject matter experts [SMEs] for others). For example, the study of dietetic practice conducted by D'Costa (1986) contained elements of functional job analysis (Fine, 1986), emphasized the social and environmental context of dietetic practice, and described practice in terms of a multifaceted model. It also specified critical scenarios, which provided the basis for test items. Other examples of integrating multiple methods into a comprehensive practice analysis questionnaire include studies of nurse anesthetists (Zaglaniczny, 1993) and psychologists (Rosenfeld et al.,

1983). Future projects might follow suit by going beyond the traditional task inventory as the basis for describing occupations and professions.

Specifically designing the data collection approach to answer the question(s) of interest is extremely important. If task importance is the most critical determinant of examination content, contributors should respond to questions about task importance. Similarly, if knowing the frequency with which a particular activity or competency is required, questions about frequency should be asked explicitly. It also is critical to design the practice analysis so that the results can readily inform the development of test specifications (Raymond, 2015). See Huff et al., this volume, for more information about designing test specifications that are used to guide examination development.

It is important for the entire process of conducting a practice analysis to be viewed as an empirical activity that must be performed carefully and systematically. Done correctly, a practice analysis is likely to be resource intensive. It also is a process that should be repeated on a regular basis (e.g., every 7–10 years) so that any changes in professional practice can be identified and examination content coverage can be revised. Though potentially daunting, the effort is critical to the credibility of the examination program. As was described earlier, the legal implications are potentially significant for a credentialing program that has not done its due diligence in aligning examination content with practice requirements.

Adapting the Assessment to Changes in the Profession's Practice

In professions such as medicine and law, the potential for frequent changes in practice necessitates ongoing attention to evaluating practice data and making necessary adjustments to test content. New science changes medicine and new judicial decisions change law. This means that in addition to periodically updating the practice analysis, item pools must be reviewed and updated frequently. In some cases, specific events may trigger such a review; for example, the change in guidelines for the use of hormone replacement therapy that occurred in the early 2000s led to a review of related items in the USMLE item pools.

In addition to changes in practice specific to test content, technology has created more pervasive changes in how professions work. An important example of one of these technological changes is the availability of electronic resources that provide easy access to a wealth of information. For example, Myers and Bashkov (2020) reported that, on average, physicians consult resources for approximately 30% of the patients they see. A central role of credentialing organizations is to steer away from questions that assess factual recall of test content and instead assess performance related to constructs that are needed for practice (e.g., judgment on case law). This has led to several professions deciding to include online resources as part of their assessments. Aligning the assessment conditions with practice activities has the potential to support the validity of the resulting inferences that are made based on the test scores. At the same time, such changes lead to a question of whether access to electronic external resources during an assessment may unintentionally change the construct that is being assessed from actual professional knowledge, skills, or abilities in a particular area to the ability to retrieve pertinent information. Lipner et al.

(2017) provided an example of research designed to answer this question. They conducted a randomized experiment using case scenarios and multiple-choice questions to assess physicians' clinical judgment skills under timed testing conditions. They found that the inclusion of an electronic resource did not adversely affect assessment performance and did not change the construct targeted by the assessment. Due to testing time limitations, the external resource functioned as a tool much like a calculator used on the SAT exams. If, however, there were no time limit, they posit that the construct might change significantly such that what was being measured was the test taker's ability to find the correct answer through information retrieval skills. This change could make it possible for a non-physician to pass the assessment, which clearly undermines test validity.

Automated Item Generation

Recent advances in testing and technology have allowed researchers to explore the use of model-based approaches to content development that will improve efficiency in the production of test material and support the design of assessments to enhance both the quality of the resulting questions (Gorin & Embretson, 2012; Leighton, 2012) and the validity of associated score interpretations. Mislevy and Haertel (2006) stated that "evidence-centered assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument an assessment is meant to embody (p. 6)"; this conceptualization meshes well with Kane's popular validity framework (2006, 2013).

Automated item generation (AIG) is an advanced test development approach that attempts to address the need for fast, efficient production of domain-specific test items by employing a generative process that uses models and computer technology to create new test items. Where traditional item development is a resource- and labor-intensive process requiring subject matter experts (SMEs) to develop individual items one at a time, AIG automates and streamlines the process by using technology to rapidly generate hundreds of test item iterations. Gierl and Lai (2015) described a model-based approach to content development using cognitive task analysis (Clark, 2014). At present, relatively little is known either about the cognitive processes used by test takers in responding to test items or about the characteristics of those items that impact their psychometric performance, although one study indicates that AIG items perform similarly to other test items (Gierl et al., 2016). Although AIG continues to be studied, the technology is not currently in widespread operational use. More research is needed to evaluate whether the anticipated benefits of AIG will be realized for the approach to be adopted more broadly.

Innovative Item Types

Since the 1960s, multiple-choice items have been the most commonly used item format in credentialing contexts (M. McDonald, 2014).³ This format was introduced in 1915 as a way to (all but) eliminate the scoring errors that were rampant in educational assessments, and it has additional benefits in that it both supports broad sampling of test content in a fixed amount of testing time and reduces the effort required for scoring.

The primary disadvantage, particularly in credentialing contexts, is that test takers select a response from a predetermined list rather than constructing a response based entirely on their own knowledge and skills. At more advanced levels of credentialing, the concern is that this does not necessarily provide clear evidence that test performance will generalize to performance in the real-world practice environment. In some contexts, this disadvantage may be trivial. If, for example, it is important for Certified Public Accounts to know/be able to identify the fraudulent activities that could be perpetrated because of a lack of effective internal controls in the revenue cycle, (a) that knowledge can be readily assessed using multiple-choice questions and (b) doing so is completely relevant to what is necessary for performance in practice. In other contexts, competent practice necessitates not only content knowledge but also demonstration of how that foundational knowledge is to be applied; in some of these situations, multiple-choice items may be less appropriate. It is these situations in which there has been a focus on developing and administering *innovative* or *alternative* item types.

Innovative (or alternative) items are those presented in a format other than the traditional multiple-choice format (Association of Test Publishers & Institute for Credentialing Excellence, 2017; Parshall et al., 2002; Wendt & Harmes, 2009). Many varieties of innovative items exist, and we provide a brief overview here both to provide insight into why these item types increasingly have been used for credentialing purposes and to explain why this trend is important and relevant in credentialing contexts. (See Huff et al. and Bennett et al., both in this volume, for further information about item types.)

The hallmark of alternative item types is that at least one of the following aspects of the items is different from traditional multiple-choice items: item presentation, the test-taker task, the response method, or item scoring (Association of Test Publishers & Institute for Credentialing Excellence, 2017). Though many types of innovative items exist, several categories tend to be used most commonly: items that include audio or video as stimulus material, multiple choice multiple response (two or more keys, often more than four options), hot spot (answer in the form of clicking on part of graphic), drag and drop/place (clicking and dragging text or graphics to match, sort, or rank), written response (an open response to a prompt such as an essay or short-answer question), and audio/video prompt (an audio or video clip is part of the item stem; Cadle & Parshall, 2015; International Test Commission & Association of Test Publishers, 2022; Parshall & Harmes, 2008).

The decision to include innovative item types should be motivated by a need and/or desire to create assessment tasks that capture important construct-relevant information about test takers, rather than simply because it is possible and exciting to develop something new and different. In credentialing contexts, where the assessment is intended to provide some level of insight about a candidate's ability to practice competently and safely, the focus may be on including more realistic assessment tasks that provide a clearer link between test performance and performance in the profession. The reason for developing these items also may be the need and/or desire to present

tasks that require the same types of cognitive processes that are necessary in practice. The typical guiding principle behind including these items in the assessment is greater fidelity; unfortunately, greater fidelity does not guarantee validity (Parshall & Harmes, 2008). When developing these novel assessment tasks, it is important to be sure that the constructs of interest are the focus of the assessment and that the item type is selected because it adds something important to the measurement of the construct. If, for example, it is important for a medical student to be able to interpret heart sounds, one complex assessment option might be a performance-based examination in which students examine standardized patients using an electronic stethoscope that presents simulated heart sounds. An alternative approach would be to incorporate audio into a text-based multiple-choice item. The latter option is a more efficient approach that likely allows for equal, if not better, measurement of the construct of interest.

Increasingly sophisticated technology permeates most aspects of daily life, and the domain of testing is no exception. A desire for increased authenticity has continued to influence the trend toward technologically innovative assessment solutions in which test takers engage with and demonstrate competence through activities that more closely approximate real-world practice. Computer simulations have gained in popularity, though the complexity, fidelity, and interactivity of these approaches can vary significantly (Association of Test Publishers & Institute for Credentialing Excellence, 2017; Margolis et al., 2002). A low-fidelity example might present four pictures of dermatologic conditions and have the test taker click on the one that is appropriately treated with cryotherapy (Association of Test Publishers & Institute for Credentialing Excellence, 2017). A high-fidelity example might be using a patient mannequin and simulated catheterization laboratory to assess interventional cardiologists' procedural skills. This type of simulation could yield information supporting the claim that candidates know how to do the procedure instead of just knowing the right thing to do without actually performing the task under pressure (Lipner et al., 2010). As noted previously, the evolving nature of professional practice has prompted innovations in assessment design. A notable example—aimed at more closely mirroring real-world conditions by allowing access to key resources during testing—is the American Board of Internal Medicine's decision to incorporate commonly used point-of-care tools into its certification exams (Lipner et al., 2017). Similarly, the Uniform Certified Public Accountant examination uses task-based simulations that typically require test takers to use authoritative literature provided in the examination (<https://www.aicpa.org/becomeacpa/cpaexam/downloadabledocuments/cpa-exam-digital-brochure.pdf>). Other examples of new innovations include drag and drop/place as well as hot spot items in the ARE exam (<https://www.ncarb.org/blog/a-deep-dive-are-5-item-types>), drag and drop/place items used by the National Council of Examiners for Engineering and Surveying (<https://ncees.org/exams/cbt/>), and the computer-based patient management simulations used in Step 3 of the USMLE (<https://www.usmle.org/step-3-test-question-formats/computer-based-case-simulations>).

Feedback

Test takers have increasingly demanded that testing organizations enhance the educational value of assessment by providing feedback about strengths and weaknesses. This is especially true for continuing certification programs where a new focus on learning and improvement is intertwined with the assessment (American Board of Medical Specialties, 2019). Metacognition has been shown to be important in self-regulated lifelong learning, from having the ability to adjust one's learning to becoming a better professional (e.g., a better clinician who makes fewer medical errors; Medina et al., 2017). Metacognition includes planning, monitoring, and evaluating one's own learning process, and the idea is to make this process more automatic so that errors are not repeated. For example, in the Doctor of Pharmacy program, a key goal of ensuring the reliability and validity of examination scores is to enhance student learning and achievement of the program goals; this is done by providing meaningful feedback for metacognitive learning after an examination (Ray et al., 2018). To enhance metacognition, the authors suggest an exercise after the test where cohort performance data by topic are shared with students and they engage in a three-step process of describing their understanding of concepts when they took the test, reviewing faculty feedback and identifying essential concepts, and reflecting on their knowledge gaps. Despite the clear value of feedback from the perspective of the test taker, testing organizations must contend with the natural tension between the risks and benefits of providing such information. Ideally, once the assessment is complete, individuals would benefit from receiving feedback in the form of the question, the correct answer, and an explanation of why the correct answer is right and the incorrect answers are wrong. From the perspective of the test taker, the benefits of receiving this information are clear. For the testing company, however, the associated risks of exposing examination content may not be acceptable.⁴

Much has been written about score reporting guidelines and feedback (Zenisky & Hambleton, 2015) and the influence of the context of the testing experience on what makes a report useful and understandable (see also Zenisky et al., this volume). Klesch (2010) complemented this literature with detailed work on score reporting for teacher certification, specifically identifying ways to use technology to provide better and more engaging feedback to teachers on their performance. She also noted that raw scores, percentage scores, or even narrative performance descriptions were more desired than scaled scores and that use of statistical terms or abbreviations was not recommended. Likewise, information about confidence intervals (i.e., measurement error) was not useful for those preparing to take the test again. Klesch found that providing information in many different ways (e.g., contextual, statistical, visual) was important for different learning styles.

Short of giving the full question/answer/rationale, some medical testing organizations have provided feedback in the form of subscores (performance scores for subdomains of an exam) to provide more diagnostic performance information (e.g., areas of relative strength and areas in need of improvement). An example of this type of feedback can be seen in the Internal Medicine Certification exam in which

subdomain scores are presented to test takers following the examination (American Board of Internal Medicine, 2023). Presenting subscores introduces an array of measurement concerns because assessments are designed to measure the overall construct of the discipline such that the total score is the most appropriate measure of the test taker's knowledge in the field (e.g., Feinberg & Jurich, 2017; Feinberg & von Davier, 2020; Feinberg & Wainer, 2014a, 2014b; Sinharay, 2010; Zapata-Rivera, 2018). Despite this caution, test takers desire performance information because it helps them to understand their areas of strength as well as areas that are in need of improvement. Particularly in credentialing contexts, test takers who are not successful often request additional information specifically to aid them in understanding where they fell short so that they can address those weaknesses in a focused way. Subscore use has increased over the years, and this makes it even more important for testing organizations to ensure that reported subscores are reliable and valid. Augmentation of subscores with information obtained from other parts of the assessment sometimes is done to achieve this goal, especially when the subscores are correlated (de la Torre & Patz, 2005; de la Torre et al., 2011; Edwards & Vevea, 2006; Sinharay et al., 2011; Wainer et al., 2001). Puhan et al. (2010) showed that reporting subscores may be most beneficial (or at least not harmful) at an institutional level (e.g., programs where 30 or more students were trained) but that they typically do not add much value over the total score. However, augmenting subscores using information from the total score did result in increased subscore reliability. Similarly, augmenting subscores using even just one prior testing occasion can improve subscore reliability, especially when the number of domains is limited and the overall test length is relatively short (Smiley, 2019). Feinberg and von Davier (2020) described a method for identifying unexpectedly high or low subscores that could stand out for test takers and be more actionable. This may help to limit misinterpretation of subscore reporting.

Another approach to enhancing the educational value of an assessment is to provide feedback at the topic and task level such that the intended measurement objectives of the incorrectly answered questions are revealed without exposing the exact item content. This enables test takers to understand their knowledge gaps at a more detailed level than does providing simple subdomain scores. The development of this approach is described in detail elsewhere (Brossman, 2018), and an example of the final presentation of the score report feedback is presented by Lipner et al. (2019). The obvious concern about this approach is low subscore reliability; there are few questions in any single subdomain because the assessment is a sampling of the entire domain. Focusing on areas with low reliability may give test takers a wrong signal of where their weaknesses lie. When there is significant variability in the number of questions represented by a subscore, there also may be instances in which studying the topic area with more questions is a better strategy than studying the topic area in which the test taker had relatively poorer performance. That said, providing more specificity does help test takers better understand what they got wrong on the test

and, as long as they are aware of the limitations, it helps them focus on the topics where test questions were missed.

A slightly different perspective on the provision of feedback relates to contexts that employ a hybrid assessment approach intended to satisfy both formative and summative assessment needs. Formative feedback is provided with the goal of helping learners identify areas of strength and areas in need of improvement so that they can focus their efforts on those areas for the summative component. This practice can lead to challenges in interpreting test results and in evaluating the overall validity of the inferences made from those scores. This topic is further discussed in the “Validity” section, specifically from the perspective of consideration of threats to test validity. In the next section, we address aspects of reliability that either are unique to or warrant special consideration in credentialing contexts.

RELIABILITY

Reliability relates to the estimation of error in measurement that might arise from different sources. The measurement community has developed several formulas and indices to estimate the error of scores, scorers, decisions, and classifications (Haertel, 2006; Lee & Harris, this volume) and, more specifically, has provided practical insight into the applications of these measures in credentialing contexts (B. E. Clauser et al., 2006). The expected reliability for licensing and certification examinations will depend on multiple factors. For example, the types of items, scoring methods, decisions, and decision rules that are used for the examination will influence the selection of methods to estimate reliability. Practitioners are encouraged to use methods that focus on the greatest potential sources of error. To illustrate this point, for a human-scored performance task that has subjective features of a rubric, one important source of error likely is in inconsistent application of scoring criteria to the performance; when the assessment is based on relatively few tasks, the sampling of tasks (the person-by-task interaction in generalizability theory terms) also may make a substantial contribution to error.

Kane (1996) discussed another concept that is particularly important to credentialing tests: error tolerance. The high-stakes nature of any test used for classification and decision-making increases the need for precision around the passing (cut) score. The conditional standard error of measurement (CSEM) around the cut score provides users with an estimate of the uncertainty of the decision: The larger the CSEM, the greater the uncertainty. The location of the cut score in the score distribution also influences the confidence and connects with Kane’s recommendation to consider the error tolerance in describing it. As users think about Type I (false-positive/pass) and Type II (false-negative/fail) errors in terms of decisions, the tolerance for these types of errors will be dependent on the risk of making each one. In the instance of a Type I error, policy makers will need to evaluate the risk to stakeholders and the public of someone entering a profession whose true score may not meet the minimum standard for competence (assuming all other eligibility and licensure or certification requirements are

met). This is where the interpretation and use of the credential is important to consider. One can make a reasoned argument that for members of the healthcare community, the risk of an unqualified candidate can have devastating, if not deadly, consequences. The risks of false-positive outcomes in other professions (e.g., architects, lawyers, educators) can be similarly costly.

Conversely, the risk of a Type II error has additional consequences, such as potentially limiting the pool of qualified practitioners available to the public. The challenge is that programs are unable to accurately determine which candidates are the results of Type I or Type II errors; it is only possible to know that there are some candidates within that range of uncertainty that may be misclassified. Because of these risks and the associated limitations, it is important for testing programs to review information about classification and discuss the levels of risk tolerance associated with different decision points. It is in these instances where the combination of validity, reliability, fairness, legal risk, and testing policy intersect.

Methods

To reduce redundancy with the discussion of methods provided by Lee and Harris (this volume), we have focused our discussion on measures of classification consistency and accuracy. The *Standards* (AERA et al., 2014) noted that licensing and certification tests should prioritize these sources of reliability evidence. It is important to note that methods for estimating reliability of scores also are applicable for credentialing testing. However, per Standard 11.14, estimates of decision consistency also should be provided, and “the consistency of decisions on whether to certify is of primary importance” (AERA et al., 2014, p. 182). As applied to licensing and certification tests, Popham and Husek (1969) discussed the issue of calculating reliability for criterion-referenced tests in an educational context when there was an expectation of limited score variability and scores well above the passing score. This led to the subsequent development of methods that could respond to estimates of classification or decision consistency (see Brennan & Kane, 1977; Brennan & Prediger, 1981; Huynh, 1976; Livingston, 1972; Subkoviak, 1976). These initial efforts represented an expansion of the concept of reliability and, by extension, an expansion of some of the evidence that would be valuable for licensing and certification programs.

An additional wave of methods emerged in the late 1980s and 1990s, driven by efforts to clarify the interpretation of these indices (S. A. Livingston, personal communication, January 18, 2002) and by advances in computing power that made it easier to implement more sophisticated techniques. Approaches described by Hanson and Brennan (1990), Breyer and Lewis (1994), and Livingston and Lewis (1995) sought to respond to these questions. Further research by Brennan and Wan (2004) described the application of bootstrap techniques. More recently, work by Cui et al. (2012) using cognitive diagnostic models provided additional value to this still growing line of research. The application of cognitive diagnostic models along with application of measurement decision theory methods (see Rudner & Gao, 2011) has seen greater utilization in licensing and certification, particularly in maintenance of certification and the use of longitudinal

assessments where test/assessment items may be used more as assessment for learning to evaluate and promote continued competence. These longitudinal approaches are currently being used (e.g., American Board of Anesthesiology, American Board of Pediatrics, American Board of Internal Medicine) and explored (e.g., National Commission for Certification of Physician Assistants) as new forms of assessment for maintenance of certification that are not event-based test-delivery experiences. In these longitudinal assessments, programs often will administer a smaller number of questions (e.g., from 1–5 to 20–30) every month or two as a strategy to facilitate ongoing engagement in professional literature and to maintain currency in the profession.

Implications for Practice

Although there is no single approach that is appropriate for all situations, it is important to understand the assumptions and limitations of available methods in order to make informed decisions about which approach to use in a given context. As it relates to the passing score, a CSEM can be interpreted, but with CSEMs often yielding relatively standard values, the use of a classification consistency index that is reported on a metric between 0 and 1 may be easier to communicate to more diverse stakeholder audiences. This means that when communicating with policy bodies or lay stakeholders, stating that a program has 90% or 95% confidence in the accuracy of its pass/fail decisions is often more interpretable than sharing a CSEM value. When these statistics are reported, it also is appropriate to note limitations or cautions with respect to the stability of the estimate. The practicality of using classification consistency estimates illustrates the interrelated nature of dependent characteristics of scores that are produced by licensing and certification exams; these include sample sizes, first-time takers versus retest candidates, location of the passing score, and related risk management considerations.

Something unique to credentialing contexts that is not frequently discussed in methods of reliability relates to smaller volume programs. The data dependency of most methods (bootstrapping or simulation being exceptions) means that estimating reliability or classification accuracy for smaller volume programs remains a meaningful challenge for which neither historical nor modern methods offer satisfactory solutions. As with the initial nudge from mastery testing within educational assessment to explore single administration estimates of decision consistency, there may be an opportunity to again look to K–12 education for consideration of additional methods or concepts that could be applicable solutions for credentialing programs.

Statistical estimates of reliability are often unstable with small sample sizes. Similarly, estimates of the conditional standard error of measurement at the passing score also may be imprecise, though they can still offer some empirical insight. However, in response to the challenge of providing evidence of reliability for the many smaller volume credentialing programs, it seems necessary to move beyond statistical calculations/estimates only. If we conceptually define reliability as the replicability of observations of samples of candidates' abilities that are judged to align with a given construct, then extending

the interpretation of reliability may be appropriate. To that point, the ideas suggested by J. K. Smith (2003) in a reconceptualization of reliability for classroom assessments may be worth exploring in greater detail. Classroom assessments often suffer from some of the same challenges as smaller volume credentialing programs: There is a small sample of test takers and the expected range of abilities for those test takers may be limited. The combination of these factors often yields unstable and potentially suppressed estimates of score reliability. Corrections for attenuation may not be enough to provide a reasonable estimate.

In rethinking the characterization of reliability, J. K. Smith (2003) suggested using a criterion of sufficiency as a proxy for traditional measures of reliability. In other words, psychometricians often recommend having a minimum number of measurement opportunities (e.g., items, questions, score points) for each important concept or domain of interest. Using that approach, and absent robust data sets to conduct traditional analyses, we may be able to make an argument that the object of measurement is being sufficiently represented by multiple observations in alignment with other sources of recommended best practices. For licensing and certification programs, the concept of sufficiency could be applied as an interim step until data can be collected to apply more traditional methods. This approach also may provide a solution for an important source of evidence for credentialing programs where the current practice may be to simply ignore consideration “due to low sample sizes.” Professional standards suggest we can do better, yet currently there is a dearth of evidence-centered approaches to guide such improvement. In 2017, the Institute for Credentialing Excellence published a revised version of the National Commission for Certifying Agencies’ Standards for the Accreditation of Certification Programs. In this document, Standard 20 addresses reliability concerns and provides specific commentary associated with estimating reliability for small-volume programs:

When candidate volumes are so small or there are other factors which lead to reliability estimates that are not meaningful, programs should describe the procedures used to demonstrate that the decisions made on the basis of scores are reasonable and fair. (Institute for Credentialing Excellence, 2014)

This comment clearly acknowledges limitations associated with providing meaningful reliability estimates in certain contexts, but limiting solutions for small-volume credentialing programs to provision of a thoughtful narrative about their procedures provides some insight into the fact that there still is a long way to go. Newer approaches, such as bootstrapping techniques, may hold some promise in advancing the assessment of reliability with small samples (Amalnerkar et al., 2020), but more work is needed to evaluate the application of these procedures to credentialing contexts.

Another unique aspect of licensing and certification tests is the opportunity for failing candidates to retest with different policy conditions governing these opportunities. One of the outcomes of this approach is that it results in inflated false-positive/passers (Type I) error rates. B. E. Clauser et al. (2006) illustrated that administration conditions with only repeat candidates lead to inflated Type I error rates. Attempts to

control this characteristic were discussed by B. E. Clauser and Nungester (2001) and included such decision rules as using average scores, most recent scores, or best scores within a defined period. Another strategy would be to potentially increase the passing score for subsequent attempts. However, this may lead to equal protection concerns if the passing expectation is differentially applied to candidates, raising questions about the extent to which any given administration is an independent observation of the candidate's ability that is not dependent on prior information or test-taking performance.

Similarly, there is an important relationship between estimates of classification consistency and the location of the passing score. This relationship extends in multiple directions. First, depending on the choice of classification consistency method, the location of the passing score relative to the distribution of candidate scores can potentially inflate or suppress the estimate. For example, this risk is highlighted if a method is chosen that is more influenced by the location of the passing score relative to the mean of the distribution. Second, as discussed in the section on legal challenges to licensing and certification tests, the location of the passing score also may be influenced by the impact on different demographic groups if there is a goal to reduce the potential effects of adverse or disparate impact.

In this section, we discussed the unique aspects of reliability as they relate to licensing and certification tests. As noted, concepts and methods are generally applicable across a range of testing applications. The interpretations and uses of scores from licensing and certification tests suggest additional and integrated considerations that are noted here as they relate to the larger validity framework articulated in the next section.

VALIDITY

Validity is a fundamental concept in educational measurement that is relevant to all aspects of the testing process, spanning test development and delivery, the analysis of testing data, and the interpretation and use of resulting scores. Lane and Marion (this volume) provide a comprehensive review of validity and validation; as such, the focus of this section will be on the aspects of validity that are most applicable to licensing and certification assessment programs. We use Kane's validity and interpretative argument as our guiding framework for this discussion.

Kane's Validity Framework/Interpretative Argument

In credentialing contexts, Kane's framework espouses employing a multifaceted approach to collecting validity evidence to maintain a program of high integrity that is fair to test takers and defensible to stakeholders. It calls for "a clear statement of the proposed interpretations and uses and a critical evaluation of the interpretations and uses" (Kane, 2013, p. 1). Four types of inferences, or claims, are important within Kane's approach: scoring, generalization, extrapolation, and interpretation/use. For assessments with scorable item types, the assigned score becomes a reflection of performance in a test setting where the assessment is taken under standardized conditions and test takers have no prior

knowledge of the specific content that is included on the test. The *scoring* component includes considerations of adherence to the defined administration procedures as well as appropriate use of raw scores (if used), item response theory, field testing, scaling, and equating. *Generalization* involves making inferences from observed scores based on the specific sample of questions on the test or the sample of observations scored by raters to the universe of questions/observations from which they were sampled. Within classical test theory, this is represented by the relationship between the observed score and the true score. Generalization is impacted by errors of measurement, which typically are random; systematic errors also may exist, and when they do, they may be both difficult to identify and damaging to score interpretations. *Extrapolation* relates to the ability to make inferences about real-world performance based on the observed scores. The fourth component of the validity argument, *interpretation/use*, refers to using the test scores to draw a conclusion about the test taker or to inform a decision such as the granting of a license.

The more ambitious the claims one seeks to make based on assessment outcomes, the more validity evidence is required. Yet the more ambitious claims typically are harder to validate because *all* inferences in the interpretive argument must hold for the interpretation to be considered valid. In licensing and certification, the concept of validity has expanded beyond the question of whether the test measures what it was intended to measure. It also emphasizes the idea that the testing organization must take broad responsibility for the consequences of the testing outcomes. A claim might be that those who pass the assessment and are deemed competent should perform competently in real-world settings. This is an ambitious claim and one that is difficult to validate, because assessments typically focus on constructs that rarely can be isolated in real life. For instance, medical assessments often test knowledge/clinical judgment in a discipline, but patient care clearly reflects more than just discipline-specific knowledge. Patient-care decisions are influenced by a myriad of factors including patient compliance, professionalism, communication skills, and systems factors such as teamwork. Although good performance on an assessment of knowledge may relate to higher quality patient care, it is difficult to control all other factors that influence patient care in a real-world setting when trying to determine the strength of that relationship. In addition, knowing the right thing to do on a multiple-choice test taken in an artificial setting is not a guarantee that the test taker would do the same thing in practice. Although some credentialing programs still use oral examinations to test practice-relevant skills, many are transitioning to sophisticated multiple-choice items that evaluate clinical judgment and the ability to synthesize information. Innovative item types also have been introduced in an attempt to address these challenges, and though initial results are encouraging, additional empirical evidence is needed to evaluate whether they are better than sophisticated multiple-choice items for assessing the desired competencies (Swanson & Hawkins, 2017). Similarly, in medical contexts, assessment approaches based on case records or logs (e.g., patient charts from electronic health records) eventually may supplement or replace multiple-choice questions as metrics based on actual practice become more universally accessible.

Importance of Extrapolation and Interpretation

All claims from Kane's framework are important in establishing the chain of evidence. For professionally developed credentialing assessments, evidence to support the scoring and generalization aspects of Kane's framework typically is much easier to produce. In most testing organizations, great effort is placed on test construction procedures: Practice analysis helps to ensure relevance to practice, standardization of testing procedures helps to ensure fairness, and scoring processes help to ensure accuracy. In addition, generalizability of test scores is demonstrated both with longer assessments covering the breadth of the field and by controlling for content and statistical specifications included in the blueprint. These are well-understood processes that have undergone significant specification and refinement over time.

Providing evidence to support the appropriateness of both *extrapolation* from test scores to real-life situations and score *interpretations/uses* is much more challenging. The high-stakes nature of licensure and certification makes the extrapolation and interpretation components critical with respect to validity. If claims are being made about the assessment that go beyond the actual test setting, an evaluation of whether these claims hold true should be done by conducting a program of research designed to generate a preponderance of evidence that either supports or disproves the claims. The higher the stakes and the claims made about the assessment results, the stronger the research program will need to be to support validity arguments. In medical licensing and certification, there is a growing body of peer-reviewed literature that provides evidence related to extrapolation to real-life situations; unfortunately, this work remains limited because definitive outcome measures are limited.

The gold standard for validity evidence in credentialing contexts is the ability to demonstrate that those who pass the assessment perform competently in practice and those who do not pass do not perform competently. In licensure contexts, however, this evidence is impossible to obtain because those who fail the examination(s) never receive a license. This type of evidence is feasible for certification exams, because certification typically is not mandated for practice. For example, a considerable amount of evidence exists to support the claim that being certified or performing better on a certification exam is linked to fewer disciplinary actions by state medical licensing boards. In internal medicine, a board-certified physician is five times less likely to have a disciplinary action than a noncertified physician (Lipner et al., 2016), and the higher the physician's score on the certification exam, the less likely they are to have a disciplinary action (Papadakis et al., 2008). This relationship has been shown in many studies comprising numerous specialties in medicine (Jones et al., 2018, 2020; Kinney et al., 2019, 2020; Kocher et al., 2008; Kopp et al., 2020; Lipner et al., 2016; F. McDonald et al., 2018; Nelson et al., 2019; Papadakis et al., 2008; Peabody et al., 2019; Zhou et al., 2017, 2018, 2019).

Other studies have sought to provide evidence to support or refute extrapolation claims by examining the relationship between measures of performance on the assessment and measures of performance in practice. One study showed that, for patients with acute myocardial infarction, treatment by a board-certified internist

or cardiologist was associated with a 19% reduction in mortality compared with treatment by a noncertified internist or cardiologist (Norcini et al., 2002). Another study showed that receiving care from an internist who scored in the top quartile of their exam was associated with a 17% increase in odds of *guideline-compliant* diabetes care compared to those who scored in the bottom quartile (Holmboe et al., 2008). A third study showed that physicians who maintain certification at 20 years after initial certification have better care management on the Healthcare Effectiveness Data and Information Set performance measures (Gray, Vandergrift, Landon, et al., 2018)—including diabetes care, mammography screening, and coronary artery disease—compared to similar physicians who did not maintain certification. In Quebec's universal healthcare system, scores achieved on licensing and certification examinations show a persistent relationship, over 4 to 7 years, with specific measures of preventive care and acute and chronic disease management in primary care practice (Tamblyn et al., 2002). Additionally, research has found a strong link between a physician's diagnostic knowledge and likelihood of patient death, emergency department visits, or hospitalizations (Gray, Vandergrift, McCoy, et al., 2021), that physicians who score higher on certification exams are less likely to prescribe inappropriate medication to geriatric patients (Vandergrift et al., 2021), and that physicians who do not keep medical knowledge current are more likely to overprescribe opioids for back pain (Gray, Vandergrift, Weng, et al., 2021).

These examples represent just a few of the studies linking performance in practice with performance on credentialing exams. This type of research is extremely challenging, and controversy exists over whether correlational studies like these are sufficient or whether studies must demonstrate a more direct link between program/exam performance and outcomes as a necessary condition for supporting the extrapolation inference. Showing a causal relationship in medical education research is extremely difficult because randomized controlled trials are rare. Other studies using quasi-experimental designs have the potential to get closer to the randomized controlled trials design; two studies that did so (possibly due to a policy change that occurred within a particular time frame) revealed that physicians required to maintain certification save the healthcare system \$167 per Medicare patient, per physician, per year without sacrificing quality (compared to those not required to maintain certification). This result translates into roughly \$5 billion per year in Medicare cost savings (Gray et al., 2014). In addition, women who had not received prior mammography screening were 8.5% more likely to get screened when seen by a general internist required to maintain certification (Gray, Vandergrift, & Lipner, 2018). These studies are closer to a randomized controlled trial design than correlational studies because they compare physicians who are very close in age and demographics and who have very similar patient populations before and after a policy change. With a preponderance of correlational studies showing positive relationships between status in a program or performance on an assessment and patient care outcomes, a convincing argument can be made that assessment performance matters to patient care. At a minimum, the

evidence is beneficial to patients who, when selecting a provider, should know which physicians “know enough” to practice in their specialty.

The other concern about procuring evidence to support or refute extrapolation claims is that, even in simulated scenarios, examinations typically test knowledge and judgment but not other important behaviors (e.g., professionalism, communication skills, or ability to function effectively within their current system) that are needed by practitioners. These other behaviors undoubtedly impact practice performance differently, such that poor performance in practice could be associated with high assessment scores. However, in the studies cited previously, this seems to be the exception and not the norm. With appropriate designs, attempting to control for many of these other behaviors is possible although far from perfect. For instance, a study can attempt to control for types of systems by identifying practice type (e.g., academic or community setting) or for communication skills using ratings provided by program directors during residency and fellowship training. Solid research designs that use matched samples, nesting of patients within physicians (generalized estimating equations), and patient risk adjustments are critical to studying the relationships between performance on examinations and practice setting outcomes.

As a final comment on correlational studies, it is worth noting that Kane (2013) does not give this type of study much consideration in his discussion of validity. He argues that demonstrating correlations with another measure simply creates the need for a new validity argument for the criterion measure.

Threats to Validity

This section will cover a variety of topics that have a potential impact on the validity of testing processes and score interpretations. These are especially important in the context of high-stakes credentialing decisions.

Test Security

Test security plays a critical role in the validity of interpretations based on test scores. For large-scale, high-stakes examinations, the ability of testing agencies to offer frequent—if not continuous—testing leads to increased item exposure. This presents a significant security challenge because item exposure can provide test takers with advanced knowledge of the questions, leading to an unfair advantage on the test and undermining the validity of score interpretations. Item exposure is an ongoing issue for high-stakes testing (Way et al., 2014), and contemporary testing innovations—including frequent administrations, larger items banks, and communication technologies—have added to the security risk (Foster, 2016).

A number of technology-based advancements that use modeling approaches to content development such as evidence-centered assessment design (Mislevy & Haertel, 2006), assessment engineering (Luecht, 2012), and AIG (e.g., Gierl & Lai, 2015) are intended to improve efficiency in item production. Several studies have demonstrated AIG’s ability to generate hundreds or thousands of test items in very short periods of time, often in just seconds (Gierl et al., 2012; Gierl & Alves, 2011, as cited in Morrison

& Embretson, 2018). As long as the generated items are psychometrically equivalent and meet other quality assurance requirements, these items can be used to populate the item bank and expand the pool of active items from which tests can be assembled. This, in turn, prevents test takers from becoming familiar with item content in advance of the examination.

Automated test assembly provides a platform to build item exposure controls into the test development process through heuristics and algorithms. This can help reduce the overexposure of individual items and promote broader utilization of the full test bank. For example, in computer adaptive testing, item exposure is a frequent problem because of the tendency to select highly discriminating items more frequently than others (Hau & Chang, 2001). Exposure controls—such as limiting the number of times any one item from a bank can be used—can mitigate this tendency. Notably, while embedded exposure controls theoretically have been available for quite some time, the technological improvements that have allowed for them to be operationalized have only been made available more recently (van der Linden & Li, 2016).

It should be noted that our comments regarding AIG reflect on the potential promise of these methods. Actual results have lagged behind that promise. There are few (if any) published results showing that these items can be used in high-stakes testing without careful selection and postproduction editing. Additionally, the available evidence suggests that it will not be practical to use these items without pretesting because evaluations of the current technology for predicting item parameters based on item characteristics—or from the parameters that have been estimated for other items from the same model—are not encouraging. Finally, there are few, if any, studies that provide evidence to support the contention that items from the same model are sufficiently different, so that advanced access to one (or more) item(s) from the model will not improve the probability of a correct answer to the next item from that same model. Nevertheless, with continued technological developments, including significant advances in machine learning and artificial intelligence, realization of the promise of some of these methods may not be far in the future.

PREKNOWLEDGE AND UNETHICAL BEHAVIOR Preknowledge of examination content is a threat to validity that can unfairly change the construct that the test is intended to measure. Many review or test-preparation courses exist to prepare test takers for credentialing examinations, and these courses may *teach to the test* (or even teach the content of the actual test) instead of focusing on the construct the test is intended to measure. If these courses truly are construct focused, there is no inherent problem; if, however, the focus is on teaching tactics to pass the test, this raises concerns about the inferences that can be made based on test performance. Medical schools and training programs capitalize on using test blueprints to guide curricula. This practice could be viewed as positive in that these curricula improve professional education by covering a complete spectrum of the field. Alternatively, if the curriculum neglects important aspects of practice that are not on the test because the focus is solely on passing the test, the approach is problematic.

As the consequences of passing an examination become more important for a test taker's career, the likelihood of test takers attempting to engage in unethical behaviors increases. Security challenges associated with high-stakes assessments have changed along with advances in technology for computer delivery and remote testing (see Shermis et al. and Bennett et al., this volume). These testing improprieties threaten the validity of the inferences that can be made based on the resulting scores (Cizek & Wollack, 2017). This is critically important in terms of being able to attest that the professional whose name is on a credential did in fact *ethically* master the test. Without this validation, the credential would hold little meaning for stakeholders—especially, and, perhaps most importantly, the public. Testing organizations may employ numerous strategies to prevent potential security threats. Test takers typically are required to sign a code of conduct that clearly specifies the policies of the examination program and the consequences for not adhering to the code of conduct. Verifying test-taker identity is a critical step to help prevent proxy test takers from subverting the system. Prevention methods in test centers have improved in that test takers often do not sit near one other and questions can be easily scrambled or different forms of the test can be used. Finally, surveillance by test proctors, video recording of test takers while testing, documentation of unscheduled break time, and records of response time data collected routinely for both live and pretest items are all part of the detection and data forensic processes.

CONTINUOUS TESTING A further development in the continuous certification arena is a testing paradigm in which test takers answer questions on a year-round basis in a combined formative/summative assessment process (Continuing Board Certification, 2019). As previously mentioned, testing organizations have been under significant pressure to loosen security measures and increase convenience and flexibility; as a result, at-home, unproctored testing at the test taker's convenience is becoming the norm. In addition, questions and answers along with rationales and references are provided immediately after the question is answered to serve formative assessment purposes. To keep the assessment secure, new questions are written and field tested every year because there is no longer the traditional concept of a large, stored item bank with questions that can be reused numerous times. Field-tested questions that survive then can be used as live scored questions. Although (generous) time limits typically are imposed per question, test delivery procedures allow access to external resources during the test. This new testing paradigm intensifies security issues, because, in addition to the demand for a large number of test questions and the test being taken at home at any time of the day, the message that the program is both for learning *and* for assessment may be confusing. Learning typically allows for working in groups or consulting with peers, but in this situation it does not; security breaches are obviously harder to prevent and detect within such a paradigm. Watermarking and encoding test material have been introduced to trace information back to any individual who tries to subvert the system. Codes of conduct need to be enforced and strengthened, and more sophisticated web-patrolling programs and data forensics techniques must be implemented.

Guidelines have been written about security measures in online testing (Foster, 2013). However, there are unique challenges associated with at-home, online, unproctored tests, and more research is needed to help make informed decisions about the most effective way(s) to incorporate these requests without increased security breaches that may go undetected and ultimately affect the validity of score interpretations. The increasing demand for more flexible and convenient test delivery models has made the move away from brick-and-mortar centers to at-home delivery a reality that will necessitate significant enhancements in prevention and detection techniques, possibly taking advantage of the explosion of artificial intelligence capabilities.

Construct-Irrelevant Variance

Unintended factors that influence test performance pose additional threats to the validity of inferences drawn from examination results. One such factor is content bias, which can advantage certain subgroups over others. Predictive bias is concerning if the exam is meant to measure a construct for the population but is found to give different predictions for subgroups of that population who should be equivalent on that construct (e.g., females and males, Asians and Caucasians). When bias exists, we are measuring construct-irrelevant variance and the test may be seen as unfair.

Use of a statistical method known as differential item functioning is a common initial analytic step for identifying biased items. After running the differential item functioning analysis, items displaying differential item functioning are presented to experts who carefully review them and determine whether the content is truly biased (e.g., Holland & Wainer, 1993). Eliminating questions containing bias is an extremely important step in maintaining the fairness of an examination. In licensing and certification, differences in performance between groups often can be explained by other relevant factors, rather than being indicative of true bias (O'Neill et al., 2022). For instance, medical oncologists who are female more often focus on breast cancer, whereas more male oncologists focus on prostate cancer. These practice tendencies may help to explain a difference in performance that is *not* considered bias.

With computer-based testing, the automatic collection of metadata such as response time and answer changes has contributed to exponential growth in the study of aberrant behaviors (e.g., Lu & Sireci, 2007; Margolis & Feinberg, 2020). Examination time limits are important from a practical, administrative viewpoint and might be important to the construct if the testing organization believes that answering a question within a realistic time frame is important. This issue becomes especially relevant when testing is in the open-book format because typically the construct is not about whether a test taker can look up the correct answer, but rather whether they can synthesize information in a timely manner to arrive at the correct answer. If a layperson can pass a licensing examination by looking up answers over an unlimited time period, the interpretation of any performance on the examination would be meaningless. Time limits are necessary in most contexts, but those that impose unrealistic constraints introduce construct-irrelevant variance into measurement of the construct.

Other Threats to Validity

Less-obvious threats to validity in the credentialing arena can arise from disciplinary pressure to create more individualized certifications—even in well-established areas—resulting in highly specialized credentials for very small populations. (Continuing Board Certification, 2019). Certificate programs have been established in disciplines that may have fewer than 50 test takers a year. The small number of test takers in these populations presents challenges in following best practices in measurement, including field-testing questions, estimating item parameters (calibration), and scoring and equating across forms (Finch & French, 2019). Sample size affects the accuracy and efficiency of model parameter estimates that are used in making decisions about individuals. When high-stakes decisions—such as admission to training programs or employment—are based on exam performance, small sample sizes can undermine the validity of the resulting inferences. In fact, *Applied Measurement in Education* devoted an entire journal issue to studies that seek solutions to the limitations associated with small sample sizes (2020, Volume 33, Issue 1).

A final threat to validity relates to situations in which credentialing programs permit test takers to narrow the focus or scope of their assessment. This approach questions the general construct that is being tested and how that credential is being represented to the public. For instance, if test takers whose primary clinical focus is breast cancer take an assessment that focuses on breast cancer but then are provided with a certificate in general medical oncology, there are consequences for patients who might not be aware of this narrower examination focus. This begs the question of who the certificate is for: the physician, the employer, or the patient population (Vandergrift et al., 2020). Confusion over what exactly the certificate means could raise threats to the validity of the inferences based on test performance. However, asking questions that are irrelevant to those physicians that narrow the scope of their practice also can be seen as a threat to validity in that it is not measuring exactly what they are doing in practice but more what the typical physician practices in the discipline. More narrow scopes and credential designations have the potential to address this controversial issue.

Validity considerations are central to any testing program. In high-stakes contexts, where credentialing examination scores influence critical decisions about individuals, collecting robust evidence to support score interpretation and use becomes even more essential. A validity argument is only as strong as the weakest link in the chain of evidence; attention to the strength of the argument is paramount.

STANDARD SETTING

In educational testing, standard setting refers to the process of determining “how much is enough” with respect to test performance (Kane, 2017). In credentialing contexts, the question more specifically becomes, What level of competence is enough to consider a candidate minimally qualified to enter practice or to use a specific practice

designation? In these contexts, standard setting begins with the assumption that any given discipline can be defined by an underlying body of knowledge, skills, and abilities that are required for practice (Way & Gialluca, 2017). Proficiency in a domain can be measured along a continuum, and there is a theoretical point at which the difference between competence and lack of competence can be distinguished. The intent of the standard-setting process is to first describe that body of knowledge, skills, and abilities and then to estimate the point along the continuum that best distinguishes between those candidates who have enough of what is needed for safe and competent practice and those who do not (Way & Gialluca, 2017).

The above logic indicates that standard setting involves specifying two types of standards: content standards and performance standards. *Content standards* are “collections of statements that describe specific learning outcomes or objectives” (Cizek, 2012); they relate to the underlying domain of knowledge, skills, and abilities in a given discipline and the need to review the domain to identify what must be mastered for a particular purpose. In credentialing, the domain review typically focuses on what is necessary to be considered competent versus not competent to receive the credential. *Performance standards* are defined by the “distinctions between two adjacent categories of performance (e.g., between acceptable performance and unacceptable performance” (Kane, 1998, p. 130). Setting performance standards therefore refers to specifying the level of performance that indicates that the content standards have been met. Once content and performance standards have been specified, there is yet another step: determining the *passing score* (the point on the score scale that corresponds to the identified performance standard; Kane, 1998). As Kane (2017) explained, a performance standard is a *minimally adequate level of performance* for the given purpose, and a passing score is a *point on a score scale* that divides test takers by the level of performance. Passing scores therefore provide an objective way to answer the central policy question about the adequacy of a given test performance.

Though standard-setting considerations are largely similar across educational and credentialing contexts, a number of areas specific to credentialing are addressed in the *Standards* (AERA et al., 2014). In the following section, we touch on some areas that we feel warrant explicit consideration in the context of credentialing examinations. For a comprehensive review of standard setting in educational and credentialing contexts, including a detailed review of standard-setting methods and activities, the interested reader is referred to Cizek (2012), Cizek and Bunch (2007), Hambleton and Pitoniak (2006), and Ferrara et al. (this volume).

Standard-Setting Methods

Standard setting for credentialing examinations typically involves using one (or more) of many judgment-based processes to collect information from content experts about the level of performance that is necessary to meet the content standards; this information is used to determine the examination score that is associated with the minimum acceptable

level of test performance (Institute for Credentialing Excellence, 2018). Perhaps the most important consideration with respect to the selection of a standard-setting methodology is the extent to which that methodology aligns with the purpose of the testing program and its associated models of achievement (Kane, 1998).

There are two overarching categories of standard-setting approaches: test centered and test taker centered. Test-centered approaches are those in which panelists review test items or tasks and, for each one, make judgments about the expected level of performance for minimally qualified test takers (those just meeting the performance standard); item-level judgments then are aggregated to produce a passing score. In test-taker-centered approaches, panelists provide judgments about test-taker performance based on an external criterion or overall test performance; the passing score is set by identifying a point on the score scale that is most consistent with the judgments (Kane, 1998; Kane et al., 1999). Test-centered approaches are used most commonly in licensing and certification; this likely is because of the preponderance of credentialing programs that use objective (rather than performance-based) tests in the credentialing process (Buckendahl & Davis-Becker, 2012; Institute for Credentialing Excellence, 2018; Kane, 1998; I. L. Smith & Springer, 2009). (To the extent that licensing and certification examinations move toward more performance-based testing and away from multiple-choice testing, this balance could shift.) Despite the existence of numerous test-centered methods that promote different approaches to data collection, relatively few are used consistently in licensing and certification, with the Angoff method and its derivative (modified Angoff) procedures being used most frequently (Buckendahl & Davis-Becker, 2012; Hurtz & Auerbach, 2003; Institute for Credentialing Excellence, 2018; Way & Gialluca, 2017).

As typically implemented, modified Angoff procedures require panelists to review test items and provide estimates of the proportion of minimally qualified candidates that would answer each item correctly; the average of the sums of the probabilities from all experts across the set of items becomes the panel's final passing score. This procedure is attractive in its simplicity and in its flexible applications, such as being able to adapt it to a variety of item types (Buckendahl & Davis-Becker, 2012; Plake & Cizek, 2012). Despite these general strengths, these procedures have several associated limitations, perhaps most notable of which is that the task (of estimating the probability of a minimally qualified candidate answering questions correctly) tends to be a difficult one. For example, panelists—particularly those for whom more time has passed since initial licensure or certification—may struggle with appropriate expectations for a minimally qualified candidate, which results in passing scores that the testing program would consider unreasonable (Professional Testing Corporation, 2019). That is, unless panelists have been licensed or certified relatively recently, they may not have a clear sense of the level of knowledge and experience of a test taker seeking the credential, and this may lead them to overestimate the performance of a minimally qualified candidate. Another concern is that panelists may not have a clear understanding of the

extent to which test takers prepare for credentialing examinations, and therefore they may underestimate the performance of the minimally qualified candidate. These concerns lead to questions about the meaningfulness of the resulting data (Busch & Jaeger, 1990; B. E. Clauser et al., 2002; B.E. Clauser, Harik et al., 2009; B. E. Clauser, Mee, et al., 2009; B. E. Clauser et al., 2013; Mee et al., 2013).

To address these concerns, supplementary methods can be used alongside the Angoff process to offer panelists alternative perspectives on this critical task. For example, using a compromise approach such as the Hofstee method (Cizek, 2006; Ferrara et al., this volume; Thompson, 2018) and presenting those results to panelists along with their Angoff judgments allows them to see the outcomes and impact of their item-level Angoff judgments (i.e., passing scores, percentage of items answered correctly, associated fail rates), compare those results with their normative expectations for fail rates and content mastery (i.e., minimum and maximum failure rates and minimum and maximum percentages of items answered correctly), and make any necessary/desired adjustments to their judgments. (If these data are not presented to panelists for review and modification, they can be shared with the decision-makers as they consider all relevant information.)

Though criterion-referenced standard-setting approaches are most appropriate for credentialing examinations (AERA et al., 2014), the reality is that incorporating some consideration of the performance of the population of test takers can provide panelists with valuable context and allow them to make more meaningful judgments (Bowers & Shindoll, 1989). The preceding example illustrates tensions that may exist between making criterion-referenced versus norm-referenced judgments and is intended to convey one aspect of the complexity associated with standard setting in credentialing contexts. There are certain standard-setting methods that may be more or less appropriate for credentialing examinations, but that does not mean that there is a single approach or procedure that is best for all credentialing contexts. The specific features of each context must be evaluated, the benefits and challenges of relevant approaches should be reviewed, and a method or combination of methods should be selected that is most likely to yield meaningful and, perhaps most important, defensible results.

Standard-Setting Activities

Content-based standard setting involves numerous steps that include, but certainly are not limited to, recruiting content experts to serve as standard-setting panelists (or judges) and provide the required judgments, facilitating the standard-setting meeting(s), and collecting or assembling data from additional sources to aid policy makers in making a passing score decision. These general activities tend to be similar across credentialing (and many other) contexts; the differences are in the details and in the way(s) that specific organizations approach and prioritize those details. In the sections that follow, we briefly describe aspects of the above activities that we believe warrant explicit consideration for credentialing examinations.

Panel Selection

Because the judgments of panelists directly affect passing scores and the overall credibility of the standard-setting process, panel composition is of critical importance to the process (Plake et al., 1991). Raymond and Reid (2001) asserted that panelists should meet two critical requirements:

1. Individually, participants should have extensive knowledge of and experience with both the subject matter and the test-taker population; and,
2. As a group, participants should be representative of all stakeholder groups.

An additional requirement, which may be assumed but should in fact be explicit, is that participants must have a commitment to protecting the security of the testing materials (Nancy Tippins, personal communication, September 10, 2023).

When identifying relevant stakeholder groups, Buckendahl and Davis-Becker (2012) argued that stakeholders differ between licensure and certification programs. In licensure testing, the primary stakeholder is the public, which is interested in protection from harm caused by unqualified practitioners. Industry professionals, who have an interest in protecting the integrity and reputation of the profession, and educators, who may use testing information to evaluate and inform curricula, instruction, and accreditation, also are relevant licensing stakeholders. Finally, though candidates themselves are stakeholders, their *opinions* are not of primary importance because they likely have a strong self-interest with respect to entering the profession. That said, the candidates' interest in ensuring that entrance into the profession is not unduly restrictive is completely legitimate and should be reflected in the final decision.

Conversely, certification typically is a voluntary pursuit where the candidate seeks to demonstrate greater knowledge, skills, or abilities than is required of entry-level practitioners (Buckendahl & Davis-Becker, 2012). Therefore, relevant stakeholders for certification exams differ from stakeholders in licensure exams. The primary stakeholders in certification may be the candidates themselves, who seek to enhance their credentials for a variety of reasons. Employers—many of whom consider certification status when determining eligibility for a position or for making hiring decisions—also may be considered a primary stakeholder in the certification process (Buckendahl & Davis-Becker, 2012). (In some professions, the public will be the employer and so will have a clear interest in the integrity of the certification.) The professional community, which has an interest in maintaining the integrity of the profession, also is likely to be an interested stakeholder. Patients sometimes are considered stakeholders, as certification can be a process to let the public know about candidates' competency in specialized areas. Finally, credential-sponsoring organizations are stakeholders in that they have an interest in promoting awareness and use of their products and/or services.

Any standard-setting process that influences public policy will inevitably impact stakeholders (Cizek, 1993; Cronbach, 1982). As such, researchers agree that participants should adequately represent stakeholders in the standard-setting process (Hambleton & Pitoniak, 2006; Loomis, 2012; Raymond & Reid, 2001). Buckendahl and

Davis-Becker (2012) explained that in the context of licensure, recently credentialed practitioners offer a thorough understanding of the content and expectations of professionals with the target credential. Experienced practitioners may have more specialized practices and so may have more advanced skills and knowledge in some areas; they also may lack proficiency in some areas in which they no longer practice. Educators have a strong knowledge of curricula and are most familiar with the target population, but they may lack a thorough understanding of what is required in practice.

Increasing stakeholder representation also can present a number of challenges. Buckendahl and Davis-Becker (2012) noted that broad panel composition may be complicated by stakeholders with competing interests. Policy decisions require the integration of content-based knowledge and stakeholder values. However, stakeholders may have differences in opinions and motivations that are not easily reconciled (Kane, 1994b). For example, experienced practitioners may desire higher passing standards if they result in a less-saturated professional market. In contrast, trainers and educators who use student performance as evidence of programmatic success may desire lower passing standards. Recently licensed professionals also may have diverse interests: Some may push for higher standards to preserve the level of rigor they were faced with; others may represent student-level groups that actively advocate for the elimination of licensure exams in favor of diploma privilege. In an effort to balance conflicts of interest, Buckendahl and Davis-Becker (2012) recommended equivalent participation by experienced practitioners, recently licensed or certified practitioners, and educators.

In credentialing examinations, there is a debate over the inclusion of panelists who lack expert-level knowledge given the focus on increased representation from diverse stakeholders. Certification and licensing exams typically cover a wide range of topics. However, it is the nature of an expert to have specialized knowledge (Raymond & Reid, 2001). Content expertise has been regarded as the most important qualification for judges (Buckendahl & Davis-Becker, 2012; Downing et al., 2006), and a lack of expertise could undermine the credibility of the judges and the soundness of the standard-setting procedure (J. C. Clauser et al., 2017; Norcini & Shea, 1997). Kane (1994b) suggested that panelists who lack the requisite expertise to knowledgeably rate all items should only respond to items that they are qualified to address. However, additional data analysis would be required to account for the missing information. There are, of course, standard-setting contexts in which important stakeholders are likely to completely lack the knowledge and experience to make the judgments required of panel members; for example, members of the public cannot make the probability judgments required of the Angoff procedure for items from a medical licensing test. In this circumstance, the interests of that stakeholder group can be incorporated into the process as part of the instructions to the panelists. In addition, in situations in which the estimated passing score from the standard-setting exercise is reviewed and approved by examination governance, the interests of otherwise unrepresented stakeholders can be reflected in their decision process.

Nonetheless, it remains critically important to consider stakeholder perspectives in the standard-setting process because they represent the interests of groups with bona fide stakes in the outcome of a testing program (Hambleton & Powell, 1983). As noted by Kane (1994b), stakeholder input is a time-honored tradition and part of the observed democratic processes for establishing public policy, thereby providing evidence of the reasonableness of policy decisions related to a particular standard.

Procedural Factors: Training and Practice

We cannot underscore enough the importance of providing ample training and practice opportunities for panelists during the content-based standard-setting activity. As previously mentioned, panelists are recruited because of their content expertise and their experience with the relevant population of test takers. There is no reason to believe that they have special skills in content-based standard setting; in fact, in some contexts there is an explicit requirement that panelists *do not* have such experience. As such, it is wise to consider the training and practice aspects of the process to be as important as, if not even more important than, the process of collecting the operational judgments.

This training and practice can come in many forms. The first relates to providing specific information about the examination of interest. A thorough review of the examination purpose and structure is a good way to ensure that all panelists are beginning the process with a reasonable understanding of the test they are about to review. Similarly, it often is beneficial to have panelists participate in a session during which they complete a set of items much like a test taker would; this provides a general sense of the test-taker experience, which is directly relevant to their standard-setting judgments. The judgment task is to make decisions about an item based on how a minimally qualified candidate *would* answer the item; focus on the *would* (rather than the *should*) represents consideration of many factors, including the reality of the testing situation. Having the panelists take the test as the test takers do, for example, including a set of questions (a) without access to the correct answers and (b) with the same time limits that are used for the operational test, can provide valuable insight into the test-taker experience, particularly for panelists who are farther away in time from the credentialing process. Engaging in several rounds of judgments has been another feature of standard-setting activities that benefits the process by allowing panelists to practice with the overall task—and its component parts—prior to providing judgments for the data set that will be used to derive the cut score. Many different approaches to this iterative process have been described in the literature, and there is no single approach that has proven to be best. Please refer to Ferrara et al. (this volume) for additional information about workshop design and execution.

A final recommendation relates to replications of the content-based standard-setting process: Programs that have the resources to do so should consider repeating the process with different panels on multiple occasions. This provides examination governance with important information about the precision of the estimated passing score.

Decision-Making: A Policy Perspective

Although standard setting is typically viewed as a statistical and procedural process, it also involves elements of policy-based judgment (Cizek, 2006). As the *Standards* noted, the process of standard setting incorporates both value judgment and empirical considerations (AERA et al., 2014). Results from content-based standard-setting panels are one of the most important factors in establishing a passing score, but that source of data is not the only one used to inform a final passing score decision. Results typically are combined with additional sources of diverse, relevant, and often policy-related information to arrive at the most appropriate recommendation for a passing score in a given credentialing context (Geisinger & McCormick, 2010).

Proposed test score interpretations are not valid unless they are supported by evidence (Kane, 1994a; Kane 1994b). That is, the standard-setting process must be supported by evidence to validate a performance standard as an interpretation of a passing score. Specifically, Kane argued that validity of a pass/fail decision depends on the plausibility of two assumptions supporting the inference that a test taker's achievement is adequate if and only if their observed score is greater than or equal to the passing score. The first assumption, termed the *descriptive assumption*, is that the passing score corresponds with a particular performance standard. The second assumption, the *policy assumption*, is that the performance standard is appropriate for the purpose of the decision. The descriptive assumption, which relates to the content-based aspect of the standard, can be empirically evaluated. The policy assumption is more complex. Policy decisions regarding "how much is enough" (e.g., how much knowledge is required to demonstrate competence) are based on assumptions about the consequences of various choices and the values they reflect. Some assumptions about the consequences of choices that frame policy decisions can be empirically confirmed, but assumptions about values necessarily imply judgment. Therefore, to validate the policy aspect of the standard-setting process, evidence must demonstrate that the passing score and associated performance standard are reasonable. The following sections briefly describe several sources of evidence that we feel are of particular importance in credentialing contexts.

STANDARD ERROR Evidence supporting the policy assumption can be drawn from a number of different sources. The standard errors of various methods, for example, can provide procedural evidence of the reasonableness of the process (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Cross et al., 1984; R. L. Smith & Smith, 1988). The standard error of a passing score refers to the likely variation in passing scores that would result if the standard-setting activity were repeated with different items and/or panels (Kane, 1994b).

PSYCHOMETRIC INFORMATION Presenting pass rates and other psychometric information can promote consistency in how judges approach the standard-setting activity and also provides useful procedural evidence supporting the validity of the

policy portion of the standard-setting process (Hambleton & Powell, 1983; Jaeger, 1989; Kane, 1994b; Linn, 1978; Shepard, 1980). Statistical data on the performance of test-taker groups, including pass rates, can aid judges in setting cut scores at realistic levels. As Loomis (2012) described, in many standard-setting methods judges make initial recommendations and are subsequently provided with estimated pass rates associated with their recommendation. They then are given the opportunity to adjust their recommendations in light of the data.

In many credentialing organizations, the results of the standard-setting exercise are not used directly to establish a cut score but instead are considered one source of data intended to inform the governance group that is responsible for establishing the final cut score. That group has the authority to accept or modify the recommended cut score, and they may adjust the recommendation to ensure that different pass rates withstand public scrutiny and meet the demands of the profession (Geisinger & McCormick, 2010). For example, a decision-making body may want to compare expected pass rates with the pass rates of previous years because a substantial disparity would undermine the public confidence in and validity of the cut score. However, decision-makers should avoid adjusting cut scores to conform to a desired passing rate, especially if doing so is contrary to the goals of the standard-setting process (AERA et al., 2014; Geisinger & McCormick, 2010). In licensing, the goal of standard setting is to determine a cut score that identifies which candidates have the competence required for safe and effective practice. As such, pass rates cannot justify alterations that move the cut score outside the range of minimal competence.

CLASSIFICATION ACCURACY Data on classification accuracy is a critical source of evidence that may be used to support the policy portion of standard setting. False positives occur when a candidate's true score is below the cut score and their observed score is at or above the cut score; false negatives occur when a candidate's true score is above the cut score and their observed score is at or below the cut score (B. E. Clauser et al., 2006). In licensing and certification, it often is the case that classification errors are not equally important. For example, passing an incompetent physician may be deemed more problematic than failing a competent one. In such situations, it is reasonable for normative data to impact the recommended or adopted standards (Hambleton & Powell, 1983).

In credentialing contexts, the individual test takers bear the cost of false negatives and society bears the cost of false positives. These costs are difficult to approximate and balance. Nonetheless, balancing classification accuracies is an important tool for policy-related decision-making because it accounts for the societal costs of various decisions (Geisinger & McCormick, 2010).

Some sources of evidence are subject to stronger critique than others. For example, while the nature of the professional market is often a supplemental factor in determining cut scores for employment testing, the use of this information in credential-

ing assessments is far more dubious. The *Standards* explicitly state that this is not appropriate in the credentialing context (AERA et al., 2014). However, Millman (1989) found that cut scores on licensing tests often were lowered in situations where the supply of professionals was not able to meet market demand (Geisinger & McCormick, 2010).

Providing opportunities to retake a failed exam also has the potential to influence policy decisions because these opportunities can change the consequence of cut scores. For example, decision-makers may raise cut scores to allow for a higher initial failure rate (Geisinger & McCormick, 2010) but a lower false-positive rate across test administrations. Some have even suggested raising the cut score for subsequent examinations for test takers who have failed in a previous administration (B. E. Clauser & Nungester, 1999; Millman, 1989). However, any policy decision that incorporates retesting as a factor should first consider whether the standard-setting panel was instructed to consider retesting opportunities; if they were, further adjustment may be inappropriate.

In credentialing, this distinction is critical because achieving an examination score above the cut score is interpreted as an indication that the test taker has the requisite knowledge to receive the credential (or is a step toward receiving the credential if multiple examinations are required). Standard setting in credentialing contexts therefore has direct and potentially significant consequences on the test taker and the public with respect to permitting or denying either entry to a profession or the ability to use a specific designation to practice.

The Role of Judgment

Standards (and cut scores) for a given examination are fundamentally the product of human judgment (Kane, 1994b) and represent the intersection between a complex set of policy decisions informed by empirical data and the interests of a diverse group of stakeholders. As such, the inherent subjectivity of the standard-setting process has led to criticism about the “arbitrariness” of cut scores (Glass, 1978). Cut scores typically result from a mathematical transformation of the judgments made by a group of standard-setting panelists onto a score scale; the resulting cut score then may be presented to the policy makers so that they can make the final decision about where on the score scale the cut score should be located. These activities represent a string of judgment-based processes that understandably might lead to the perceived arbitrary nature of cut scores: panelists make judgments about test material or test-taker performance, the recommendation of a single cut score based on the panelists’ judgments often proposes only one score from a range of legitimate scores, and the credentialing bodies have the authority to use their judgment to accept or modify cut score recommendations (Kane, 1994b). The inclusion of human judgment is unavoidable, however, because there is no statistical answer to the question, How much is enough? (Kane, 1998, 2002). Although aspects of the process may be perceived as arbitrary because of its foundations in human judgment, it is important to note that an arbitrary standard is not necessarily a capricious standard (Hambleton, 1978; Hambleton & Pitoniak,

2006). Capricious standards may undermine the validity and integrity of the process (Geisinger, 1991; Kane, 1994b), but arbitrary standards may escape the perception of capriciousness if they demonstrate that the standards align with the goals of the testing program. This demonstration requires methodical design and documentation of standard-setting processes that yield cut scores with a data-informed, empirical basis (AERA et al., 2014; Kane, 2017; Hambleton & Powell, 1983). The importance of this process is perhaps most critical in credentialing contexts where the risks of making decisions that are *not* based on sound methodology have implications for the safety of the public and the careers of the candidates.

As such, it is important for practitioners involved in standard-setting activities to (a) be aware of the myriad factors that have the potential to influence standard-setting judgments and (b) do their best to mitigate the associated risks. Ferrara et al. (this volume) provide a detailed description of hypotheses about factors that influence standard-setting judgments. These include task features, cognitive processes, judgmental heuristics, and cognitive biases. We strongly recommend review and consideration of this information for those seeking to implement standard setting in credentialing contexts. The authors also present a framework to guide both the design of standard-setting workshops and a structured approach to gathering evidence to support the validity of cut score interpretations (see Ferrara et al., Table 12.3, this volume). Careful attention to the guidance provided here and by Ferrara et al. will help ensure that standard setting for credentialing examinations is intentional, aligned with best practices, and results in defensible cut scores.

Major Influences on the Practice of Assessment for Licensing and Certification

Substantive changes in assessment for licensing and certification have been influenced by four markers. These include: (a) the rapid evolution of information and advances of technology and tools including more sophisticated psychometric approaches (Norcini et al., 2013); (b) the rise of competency-based education (Hauer et al., 2016); (c) an increased focus on the need for integrating learning and assessment over a professional's career (Baron, 2016; Continuing Board Certification, 2019); and (d) the rise of antitesting and antiregulation sentiments (Flier et al., 2016; Teirstein, 2015).

Rapid Evolution of Information and Advances of Technology and Tools

Recent decades have seen an explosion of information and, though keeping pace with it is essential, the task is increasingly challenging. Advances in cognitive psychology have deepened our understanding of how people use knowledge and information. At the same time, technological innovations are transforming assessment practices; for example, the Dental Interactive Simulation Corporation is doing important work using simulations to assess problem solving in dental hygiene (Williamson et al., 2006).

The challenge is making sense of the complex data resulting from this type of novel assessment format.

Technological advances abound, many of which directly support healthcare professions: these include the mandate of electronic health systems for all physician practices in the United States, increased use of point-of-care resources, precision medicine, the genome project, health wearables, 3D printing, robotic surgery, and decision support systems (Waldren et al., 2017). New assessment approaches are trying to incorporate some of these advances into the testing experience to make it more authentic to actual practice, such as allowing access to online external resources (e.g., UpToDate, an electronic clinical tool) during the testing experience and using simulation technology to assess performance more directly (Levine et al., 2012; Lipner et al., 2017). However, it is important to balance the cost and feasibility of these approaches against the additional value they provide in evaluating the desired construct(s). As mentioned previously, the technology itself inadvertently may introduce construct-irrelevant variance into the measurement. As such, design and impact of the new assessment should be researched thoroughly before it is used for making high-stakes decisions about test takers.

Item response theory (IRT) has become the standard psychometric model used in large-scale testing programs (see Cai et al., this volume, for discussion of IRT and other models). Its use has expanded in recent years in part due to sophisticated technology like computerized adaptive testing, as well as because of extensions of IRT for handling polytomous items and measuring multiple traits. There also has been increased interest in assessments that reflect how people practice their work. This has led to the development of more complex assessments that use advanced psychometric techniques such as computational psychometrics to assess both cognitive and noncognitive skills (von Davier, 2017). Computational psychometrics is quite sophisticated, combining mathematical models, machine learning, and data mining approaches as well as theory-driven psychometric approaches like evidence-centered design to measure latent abilities. The psychometric models that deal with complex data generally are IRT and Bayesian networks (Levy & Mislevy, 2016). These more complex models are important in credentialing settings because they allow for emulating real-world situations, such as using the Internal Revenue Service tax code on the Association of International Certified Public Accountants (AICPA) exams. Psychometric approaches also have become more sophisticated, particularly in areas such as data forensics. This field involves statistical analysis of the rich information automatically collected during computer-based testing—such as test taker responses, keystroke logs, and timing data—and prior performance to detect aberrant behavior and test fraud (de Klerk et al., 2019).

Rise of Competency-Based Education

In 1999, the Accreditation Council for Graduate Medical Education and the ABMS introduced six core competencies to define the relevant skills, knowledge, and attitudes that a trainee or physician should possess to provide quality health care

(Accreditation Council for Graduate Medical Education, 2020). Soon after, the competency-based education movement arose in residency and fellowship training programs; this movement resulted in defining subcompetencies and milestones for each of the six competencies to track a trainee's developmental progress within each competency and provide a structured approach to providing developmental feedback (Nasca et al., 2012). Instead of the traditional paradigm in which time spent in a training program is the primary indicator of a trainee's progression, mastery of knowledge and skills is the key to assessing progression. An outcomes-based approach is used to evaluate a trainee's progress in the medical program using the milestones framework. Assessment is both formative (giving feedback to enhance learning) and summative (making a decision about whether a trainee has achieved a certain level of competence). Research on the validity of the inferences based on milestone ratings is growing in the medical community, and undoubtedly the milestones will continue to be refined as the community learns more about the effectiveness of the framework (Hauer et al., 2018).

With the rise of competency-based education, the concept of programmatic assessment also has evolved. This approach combines various assessment instruments to provide a more holistic evaluation of a trainee (Schuwirth & van der Vleuten, 2012; Uijtdehaage & Schuwirth, 2018). The notion is based on the fact that each assessment instrument has its own inherent advantages and disadvantages, and it focuses on the meaningful triangulation across instruments, the stakes involved in the decision-making, and how the validity evidence is collected as the keys to success. The approach appears to be most beneficial in controlled environments such as training programs where it is largely associated with lower-stakes use cases (i.e., in the assessment *for* learning paradigm). It is less clear how this approach could be implemented for professionals in practice where there are few controlled environments. Furthermore, combining programs of assessment *for* and *of* learning (formative and summative) must be done with careful planning, especially with respect to validity considerations. Although there is appeal to approaching individual assessment from a programmatic perspective, and there is some evidence that it can work, more research is needed (Bok et al., 2018).

Increased Focus on Integrating Learning and Assessment Over a Professional's Career

One of the most significant changes in assessment relates to efforts to combine different assessment methods (e.g., formative and summative) into a cohesive, longitudinal program that supports a professional's development throughout their career. Much controversy has existed around programs for recertification (i.e., the maintenance of certification and continuing certification for medical professionals). As a result, a commission was formed that included public members as well as representatives from physician-led organizations. *The Continuing Board Certification: Vision for the Future Commission Final Report* (American Board of Medical Specialties, 2019),

which resulted from this effort, suggests that continuing certification programs for physicians should emphasize ongoing professionalism, focusing on learning and improvement in addition to assessment of current competence, with the ultimate goal of providing high-quality patient care. Initial certification programs, representing a culmination of training and traditional point-in-time assessments, sufficed for this purpose. Maintenance of certification, however, should be more continuous in nature. The American Board of Anesthesiology was the first of the ABMS boards to restructure its maintenance of certification program to a longitudinal program, called MOCA Minute, which claims to enable physicians to continuously assess themselves and identify knowledge gaps using an online portal (American Board of Anesthesiology, n.d.). Other ABMS boards soon followed. This is not just the case in medicine: Nursing and physician assistant credentials all have certification renewal programs (American Nursing Credentialing Center, n.d.; National Commission on Certification of Physician Assistants, n.d.).

From a measurement perspective, the challenge for these programs is being able to a) deliver enough quality questions in a nonsecure environment and to b) provide good educational feedback that includes identifying and remediating areas of weakness while c) maintaining a reliable and valid summative decision after 5 years (i.e., the time span for most of the medical education programs). There are many obstacles to doing this well; some of these include threats to security (because the questions and answers typically are exposed immediately after a question is answered), the difficulty of field testing enough questions with small sample sizes, maintaining the same standard over a period of time (e.g., 5 years), and choosing a scoring model that can adjust over time to allow for improvement. Strong research agendas are critical to evaluate the success of these longitudinal programs in terms of acceptability, feasibility, and meaningful learning and measurement.

The Rise of Antitesting and Antiregulation Sentiments

There has been a growing sentiment against testing and regulation across all of testing. For instance, in 2020 the University of California system announced that the SAT and ACT would no longer be required for undergraduate admissions (Hubler, 2021; see also Camara et al., this volume). For licensure and certification, most of the controversy has centered on healthcare, particularly regarding continuing certification programs and the Step 2 Clinical Skills examination for U.S. medical students (Flier et al., 2016; Teirstein, 2015; Teirstein & Topol, 2015). In 2018, the Institute for Credentialing Excellence, a leading developer for standards for certification programs, and the American Society of Association Executives formed a Professional Certification Coalition (Professional Certification Coalition, 2023). This coalition aims to counteract the negative impact of attempts to enact state legislation that undermines certification programs designed to provide data-driven, external validation of practitioners' competence by nongovernment private certification organizations. This legislation is directly working to weaken professional standards in the healthcare industry.

Weakened criteria for certificates could have serious consequences for patient care, particularly if the certificate is meaningless due to being obtainable solely through monetary exchange. Although the issue of weakened credentialing criteria is not the same as the proliferation of fake organizations offering phony certificates (Marcus, 2021), it still threaten the value of programs that maintain defensible standards and have conducted legitimate research to demonstrate that the assessments are meaningful and relevant.

CONCLUSION

Assessment for licensing and certification represents a segment of testing focused on protecting the public from unqualified practitioners. As this chapter has demonstrated, though assessment in this context shares similarities with other testing contexts, it is the differences that set it apart from tests used for other purposes. These differences are the focus of this chapter.

From a practical standpoint, the primary issues for credentialing bodies relate to test development and measurement considerations that impact the resulting examination and, by extension, examination outcomes. Validity evidence for the procedures used to develop, administer, score, and set passing scores for the test is central to the inferences made about candidates based on the test results. The stakes associated with licensing and certification examinations—including the risks of granting credentials to unqualified candidates and creating barriers to employment, thereby limiting the public's access to practitioners—make these considerations even more critical.

ACKNOWLEDGMENTS

First and foremost, thank you to our esteemed editors Linda Cook and Mary Pitoniak for inviting us to contribute to this prestigious volume and for providing unwavering support and patience throughout the entire production process. Even a global pandemic could not stop this dynamic duo! Thank you to our reviewer collaborators, Dave Swanson, Leslie Keng, and Nancy Tippins, for their insightful recommendations that yielded a much more concise (believe it or not!) and informative chapter. Finally, thank you to Benjamin Chesluk, Brian Clauser, Joanna Gorin, Michelle Johnson, Maxamillia Moroni, and Sidhartha Reddy, whose hard work and thoughtful contributions to our manuscript were invaluable!

REFERENCES

- Accreditation Council for Graduate Medical Education. (2020). *The milestones guidebook*. <https://www.acgme.org/globalassets/milestonesguidebook.pdf>
- Amalnerkar, E., Lee, T. H., & Lim, W. (2020). Reliability analysis using bootstrap information criterion for small sample size response functions. *Structural Multidisciplinary Optimization*, 62, 2901–2913.

- American Board of Anesthesiology. (n.d.). *MOCA minute*. Retrieved October 1, 2023, from <https://www.theaba.org/maintain-certification/moca-minute/>
- American Board of Internal Medicine. (2023). *Internal medicine blueprint*. <https://www.abim.org/~media/ABIM%20Public/Files/pdf/exam-blueprints/certification/internal-medicine.pdf>
- American Board of Medical Specialties. (2019). *The continuing board certification: Vision for the Future Commission*. Retrieved June 9, 2025 from https://www.abms.org/wp-content/uploads/2020/11/commission_final_report_20190212.pdf
- American Educational Research Association, American Psychological Association, & National Council on Educational Measurement. (2014). *Standards for educational and psychological testing*.
- American Nursing Credentialing Center. (n.d.). *Renew your certification*. Retrieved October 1, 2023, from <https://www.nursingworld.org/certification/renewals>
- Americans With Disabilities Act of 1990, 42 U.S.C. § 12101 *et seq.*
- Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 36(1), 45–50.
- Association of Test Publishers and the Institute for Credentialing Excellence. (2017). *Innovative item types: A white paper and portfolio*. Beyond MCQ: A Showcase of Examples. <https://atpu.memberclicks.net/assets/innovative%20item%20types%20w.%20appendix%20copy.pdf>
- Balthazard, C. (2017, September 26). #39 *Defining professional regulation*. <https://www.linkedin.com/pulse/39-defining-professional-regulation-claude/>
- Baron, R. J., & Braddock, C. H., III. (2016, December 29). Knowing what we don't know—Improving maintenance of certification. *New England Journal of Medicine*, 375(26), 2516–2517. <https://doi.org/10.1056/nejmp1612106>
- Bok, H. G. J., de Jong, L. H., O'Neill, T., Maxey, C., & Hecker, K. G. (2018). Validity evidence for programmatic assessment in competency-based education. *Perspectives on Medical Education*, 7(6), 350–351.
- Bowers, J. J., & Shindoll, R. R. (1989). *A comparison of the Angoff, Beuk, and Hofstee methods for setting a passing score* (ACT Research Report Series 89-2). ACT.
- Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277–289.
- Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, 4(2), 219–240.
- Brennan, R. L., & Prediger, D. J. (1981) Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699.
- Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). University of Iowa.
- Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94-39). ETS.

- Brossman, B. (2018, April 12–16). *Score reports: A collaborative design between measurement, communications, and subject matter experts* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.
- Buckendahl, C. W., & Davis-Becker, S. L. (2012). Setting passing standards for credentialing programs. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 485–502). Routledge.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27, 145–163.
- Cadle, A., & Parshall, C. G. (2015). *Innovative item types: Strengths and weaknesses*. <http://www.proftesting.com/blog/2015/05/11/2015511innovative-item-types-strengths-and-weakness/>
- Chinn, R., & Hertz, N. (2010). *Job analysis: A guide for credentialing organizations* (CLEAR Research Brief). CLEAR.
- Civil Rights Act of 1964, Pub. L. 88-352, 78 Stat. 241 (1964).
- Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106.
- Cizek, G. J. (2006). Standard setting. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 225–260). Routledge.
- Cizek, G. J. (2012). An introduction to contemporary standard setting: Concepts, characteristics, and contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications.
- Cizek, G. J., Germuth, A. A., & Schmid, L. A. (2011). *A checklist for evaluating credentialing testing programs*. The Evaluation Center, Western Michigan University. <http://www.wmich.edu/evalctr/checklists/>
- Cizek, G. J., & Wollack, J. A. (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge.
- Clark, R. E. (2014). Cognitive task analysis for expert-based instruction in healthcare. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 541–551). Springer.
- Clauser, A. L., & Raymond, M. (2017). Specifying the content of credentialing examinations. In S. Davis-Becker & C. W. Buckendahl (Eds.), *Testing in the professions: Credentialing policy and practice* (pp. 64–84). Routledge.
- Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education*, 22(1), 1–21.
- Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–732). Praeger.

- Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, 46(4), 390–407.
- Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing*, 13, 65–85.
- Clauser, B. E., & Nungester, R. J. (1999). Considerations in adjusting cut-scores for certification and licensure decisions. *CLEAR Exam Review*, 10(2), 18–23.
- Clauser, B. E., & Nungester, R. J. (2001). Classification accuracy for tests that allow re-takes. *Academic Medicine*, 76(Suppl. 10), S108–S110.
- Clauser, B. E., Swanson, D. B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement*, 39, 269–290.
- Clauser, J. C., Hambleton, R. K., & Baldwin, P. (2017). The effect of rating unfamiliar items on Angoff passing scores. *Educational and Psychological Measurement*, 77(6), 901–916. <https://doi.org/10.1177/001316441667098>
- Copyright Act, 17 U.S.C. § 101 *et seq.* (1976).
- Cox v. Ala. State Bar, 330 F. Supp.2d 1265 (M.D. Ala. 2004).
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. Jossey–Bass.
- Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods of establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement*, 21(2), 113–129.
- Cui, Y., Gierl, M. J., & Chang, H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49(1), 19–38.
- Currier vs NBME, Case No. 07-J-434 (Mass. Ct. App. 2007), *vac. & rem.*, 965 N.E.2d 829 (2012).
- D'Costa, A. G. (1986). The validity of credentialing examinations. *Evaluation & the Health Professions*, 9(2), 137–169. <https://doi.org/10.1177/016327878600900202>
- de Klerk, S., van Noord, S., & van Ommering, C. J. (2019). The theory and practice of educational data forensics. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 381–399). Springer Cham.
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30(3), 295–311.
- de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, 35(4), 296–316.
- DeMers, S. T., Webb, C., & Horn, J. B. (2014). Psychology licensure and credentialing in the United States and Canada. In W. B. Johnson & N. J. Kaslow (Eds.), *The Oxford*

- handbook of education and training in professional psychology* (pp. 201–213). Oxford University Press.
- Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Research methodology: Procedures for establishing defensible absolute passing scores on performance examinations in health professions. *Teaching and Learning in Medicine*, 18(1), 50–57.
- Drew, J. (2016, April 4). NC may eliminate licensing for a dozen professions. *The Charlotte Observer*. <http://www.charlotteobserver.com/news/politics-government/article69815452.html>
- Dubois, P. H. (1970). *A history of psychological testing*. Allyn & Bacon.
- Edwards, M. C., & Vevea, J. K. (2006). An empirical Bayes approach to subscore augmentation. *Journal of Educational and Behavioral Statistics*, 31(3), 241–245.
- Elchert, D. M. (2016). *Educating students about professional licensure in health service psychology*. <https://teachpsych.org/resources/Documents/otrp/resources/Educating%20Students.pdf>. University of Iowa: Office of Teacher Resources in Psychology.
- ETS v. Hildebrant, 923 A.2d 34 (Md. 2007).
- Equal Employment Opportunity Commission, Uniform guidelines on employee selection procedures (Title VII Regulations), 29 C.F.R. § 1607.2B (1985).
- Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Journal of Educational Measurement*, 36(1), 5–13.
- Feinberg, R. A., & von Davier, M. (2020). Conditional subscore reporting using iterated discrete convolutions. *Journal of Educational and Behavioral Statistics*, 45(5), 515–533.
- Feinberg, R. A., & Wainer, H. (2014a). A simple equation to predict a subscore's value. *Journal of Educational Measurement*, 33, 55–56.
- Feinberg, R. A., & Wainer, H. (2014b). When can we improve subscores by making them shorter? The case against subscores with overlapping items. *Journal of Educational Measurement*, 33, 47–54.
- Finch, H., & French, B. F. (2019). A comparison of estimation techniques for IRT models with small samples. *Applied Measurement in Education*, 32(2), 77–96.
- Fine, S. A. (1986). Job analysis. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 53–81). Johns Hopkins University Press.
- Flier, L. A., Henderson, C. R., & Treasure, C. L. (2016). Time to eliminate the Step 2 Clinical Skills Examination for US medical graduates. *JAMA Internal Medicine*, 176(9):1245–1246. doi:10.1001/jamainternmed.2016.3753
- Foster, D. (2013). Security issues in technology-based testing. In J. Wollack & J. Fremer (Eds.), *Handbook of test security* (pp. 39–83). Routledge.
- Foster, D. (2016). Testing technology and its effects on test security. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 235–253). Routledge.
- Geisinger, K. F. (1991). Using standard-setting data to establish cut off scores. *Educational Measurement: Issues and Practice*, 10(2), 17–22.

- Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, 29(1), 38–44.
- Gierl, M. J., & Lai, H. (2015). Automatic item generation. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 410–430). Routledge.
- Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8), 757–765.
- Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A., & De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education*, 29(3), 196–210.
- Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), 237–261.
- Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 23–35). Routledge.
- Gray, B. M., Vandergrift, J. L., Johnston, M. M., Reschovsky, J. D., Lynn, L. A., Holmboe, E. S., McCullough, J. S., & Lipner, R. S. (2014). Association between imposition of a maintenance of certification requirement and ambulatory care–sensitive hospitalizations and health care costs. *Journal of the American Medical Association*, 312(22), 2348–2357.
- Gray, B. M., Vandergrift, J. L., Landon, B., Reschovsky, J. D., & Lipner, R. S. (2018). Association between American Board of Internal Medicine maintenance of certification status and performance on a set of healthcare effectiveness data and information set process measures. *Annals of Internal Medicine*, 169(2), 97–105.
- Gray, B. M., Vandergrift, J. L., & Lipner, R. S. (2018). Association between the American Board of Internal Medicine's general internist's maintenance of certification requirement and mammography screening for Medicare beneficiaries. *Women's Health Issues*, 28(1), 35–41.
- Gray, B. M., Vandergrift, J. L., McCoy, R. G., Lipner, R. S., & Landon, B. E. (2021). Association between primary care physician diagnostic knowledge and death, hospitalization and emergency department visits following an outpatient visit at risk for diagnostic error: A retrospective cohort study using Medicare claims. *BMJ Open*, 11(4), Article e041817.
- Gray, B. M., Vandergrift, J. L., Weng, W., Lipner, R. S., & Barnett, M. L. (2021). Clinical knowledge and trends in physicians' prescribing of opioids for new onset back pain, 2009–2017. *JAMA Network Open*, 4(7), Article e2115328.
- Gulino v. Board of Education of the City School District of the City of New York. (2015a, August 7), U.S. District Court, S.D. New York.
- Gulino v. Board of Education of the City School District of the City of New York. (2015b, June 5), U.S. District Court, S.D. New York.
- Gulino v. Board of Education of the City School District of the City of New York, 236 F. Supp. 2d 314 (S.D. N.Y. 2002), *aff'd in part, rev'd in part*, 460 F.3d 361 (2nd Cir. 2006), *on rem.*, Opinion & Order, Case 1:96-cv-08414-KMW (S.D. N.Y. 2012), *aff'd*, Summary Order, No. 1301001-cv (2nd Cir. 2014).

- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Praeger.
- Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 15(4), 277–290.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Praeger.
- Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. *Evaluation & the Health Professions*, 6(1), 3–24.
- Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27, 345–359.
- Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, 38(3), 249–266.
- Hauer, K. E., Vandergrift, J., Hess, B., Lipner, R. S., Holmboe, E. S., Hood, S., Iobst, W., Hamstra, S. J., & McDonald, F. S. (2016). Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among U.S. internal medicine residents, 2013–2014. *Journal of the American Medical Association*, 316(21), 2253–2262.
- Hauer, K. E., Vandergrift, J., Lipner, R. S., Holmboe, E. S., Hood, S., & McDonald, F. S. (2018). National internal medicine milestone ratings: Validity evidence from longitudinal three-year follow-up. *Academic Medicine*, 93(8), 1189–1204.
- Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum Associates.
- Holmboe, E. S., Wang, Y., Meehan, T. P., Tate, J. P., Ho, S., Starkey, K. S., & Lipner, R. S. (2008). Association between maintenance of certification examination scores and quality of care for Medicare beneficiaries. *Archives of Internal Medicine*, 168(13), 1396–1403.
- Hubler, S. (2021, May 15). University of California will end use of SAT and ACT in admissions. *The New York Times*. <https://www.nytimes.com/2020/05/21/us/university-california-sat-act.html>
- Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, 63(4), 584–601. <https://doi.org/10.1177/0013164403251284>
- Huynh, H. (1976). On the reliability of decisions in domain-referenced tests. *Journal of Educational Measurement*, 13, 253–264.
- Institute for Credentialing Excellence. (2014). *National Commission for Certifying Agencies Standards for the Accreditation of Certification Programs*. https://www.credentialingexcellence.org/Portals/0/NCCA%20Standards%202021%20DRAFT%20REVISIONS_Sept%202021.pdf
- Institute for Credentialing Excellence. (2018). *Standard setting overview for credentialing programs*. <https://www.credentialingexcellence.org/blog/rsrch18-standard-setting-overview-for-credentialing-program>

- International Test Commission & Association of Test Publishers. (2022). *Guidelines for technology-based assessment*.
- Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). Macmillan.
- Jones, A. T., Kopp, J. P., & Malangoni, M. A. (2018). Association between maintaining certification in general surgery and loss-of-license actions. *JAMA*, 320(11), 1195–1196.
- Jones, A. T., Kopp, J. P., & Malangoni, M. A. (2020). Recertification exam performance in general surgery is associated with subsequent loss of license actions. *Annals of Surgery*, 276(6), 1020–1024.
- Kane, M. T. (1994a). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions*, 17, 133–159.
- Kane, M. (1994b). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425–461.
- Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, 9, 355–379.
- Kane, M. T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education*, 10, 5–18.
- Kane, M. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment*, 5(3), 129–145.
- Kane, M. T. (2002). Practice-based standard setting. *The Bar Examiner*, <https://thebarexaminer.ncbex.org/wp-content/uploads/2018/10/710302-kane.pdf>
- Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Kane, M. T. (2017). Using empirical results to validate performance standards. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 11–30). Springer.
- Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education*, 4, 195–207.
- Kearney, M. S., Hershbein, B., Boddy, D., Jácome, E., & Nantz, G. (2015). *Three targeted approaches to expand employment opportunities*. Brookings. https://www.brookings.edu/wp-content/uploads/2016/06/THP_three_approaches_expand_employment_framing.pdf.
- Kelly v. W.Va. Bd. of Law Exam'rs, Case No. 2:08-00933 (S.D. W.Va. 2010).
- Kinney, C. L., Raddatz, M. M., Sliwa, J. A., Clark, G. S., & Robinson, L. R. (2019). Does performance on the American Board of Physical Medicine and Rehabilitation initial certification examinations predict future physician disciplinary actions? *American Journal of Physical Medicine & Rehabilitation*, 98(12), 1079–1083.
- Kinney, C. L., Raddatz, M. M., Sliwa, J. A., Driscoll, S. W., & Robinson, L. R. (2020). Association of participation in the American Board of Physical Medicine

- and Rehabilitation Maintenance of Certification program and physician disciplinary actions. *American Journal of Physical Medicine & Rehabilitation*, 99(4), 325–329.
- Klesch, H. S. (2010). *Score reporting in teacher certification testing: A review, design, and interview/focus group study* (Publication No. AAI3409610) [Doctoral dissertation, University of Massachusetts Amherst]. Scholarworks. <https://scholarworks.umass.edu/dissertations/AAI3409610>.
- Knapp, J., Anderson, L., & Wild, C. (Eds.), (2015). *Certification: The ICE handbook* (2nd ed.). Institute for Credentialing Excellence.
- Kocher, M. S., Dichtel, L., Kasser, J. R., Gebhardt, M. C., & Katz, J. N. (2008). Orthopedic Board certification and physician performance: An analysis of medical malpractice, hospital disciplinary action, and state medical board disciplinary action rates. *The American Journal of Orthopedics*, 37(2), 73–75.
- Kopp, J. P., Ibáñez, B., Jones, A. T., Pei, X., Young, A., Arnhart, K., Rizzo, A., & Buyske, J. (2020). Association between American Board of Surgery initial certification and risk of receiving severe disciplinary actions against medical licenses. *JAMA Surgery*, 155(5), Article e200093.
- LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation & the Health Professions*, 17(2), 178–197. <https://doi.org/10.1177/016327879401700204>
- LaDuca, A., Downing, S., & Henzel, T. (1995). Systematic item writing and test construction. In J. C. Impara, (Ed.), *Licensure testing: Purposes, procedures and practices* (pp. 117–148). Buros Institute of Mental Measurements.
- Leighton, J. P. (2012). Learning sciences, cognitive models, and automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation* (pp. 23–35). Routledge.
- Levine, A. I., Schwartz, A. D., Bryson, E. O., & DeMaria, S., Jr. (2012). Role of simulation in U.S. physician licensure and certification. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, 79(1), 140–153.
- Levy, R. & Mislevy, R. (2016). *Bayesian psychometric modeling*. Chapman and Hall/CRC.
- Linn, R. L. (1978). Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement*, 15(4), 301–308.
- Lipner, R. S., Brossman, B. G., & Grosso, L. J. (2019). Changes to the American Board of Internal Medicine's maintenance of certification program: Preserving core assessment values. *Journal of Applied Testing Technology*, 20(S2), 11–20.
- Lipner, R. S., Brossman, B. G., Samonte, K. M., & Durning, S. J. (2017). Effect of access to an electronic medical resource on performance characteristics of a certification examination. *Annals of Internal Medicine*, 167(5), 302–310.
- Lipner, R. S., Messenger, J. C., Kangilaski, R., Baim, D. S., Holmes, D. R., Jr., Williams, D. O., & King, S. B., III. (2010). A technical and cognitive skills evaluation of performance in interventional cardiology procedures using medical simulation. *Simulation in Healthcare*, 5(2), 65–74.

- Lipner, R. S., Young, A., Chaudhry, H. J., Duhigg, L. M., & Papadakis, M. A. (2016). Specialty certification status, performance ratings, and disciplinary actions of internal medicine residents. *Academic Medicine*, 91(3), 376–381.
- Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 9, 13–26.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–197.
- Loomis, S. C. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed.). Routledge.
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29–37.
- Luecht, R. M. (2012). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–76). Routledge.
- Marcus, J. (2021, December 26). Growing “maze” of education credentials is confusing consumers, employees. *Washington Post*.
- Margolis, M. J., Clauser, B. E., Harik, P., & Guernsey, M. J. (2002). Examining subgroup differences on the computer-based case-simulation component of USMLE Step 3. *Academic Medicine*, 77(10), S83–S85.
- Margolis, M. J., & Feinberg, R. (Eds.) (2020). *Integrating timing considerations to improve testing practice*. Routledge.
- McDonald, F., Duhigg, L., Arnold, G., Hafer, R., & Lipner, R. L. (2018). The American Board of Internal Medicine maintenance of certification examination and state medical board disciplinary actions: A population cohort study. *Journal of General Internal Medicine*, 33(8), 1292–1298.
- McDonald, M. (2014). *The nurse educator's guide to assessing learning outcomes* (3rd ed.). Jones & Bartlett Publishers.
- Medina, M. S., Castleberry, A. N., & Persky, A. M. (2017). Strategies for improving learner metacognition in health professional education. *American Journal of Pharmaceutical Education*, 81(4), Article 78.
- Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice*, 32(3), 27–35.
- Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher*, 18(6), 5–9.
- Mislevy R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues & Practice*, 25(4), 6–20.
- Morath, E. (2015, November 14–15). License law is nixed in D.C. *Wall Street Journal*.
- Morrison, K. M., & Embretson, S. (2018). Item generation. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (Vol. 1, pp. 75–94). Wiley.

- Myers, A. J., & Bashkov, B. M. (2020). *Evaluating use of an online open-book resource in a high-stakes credentialing exam* [Unpublished manuscript]. American Board of Internal Medicine.
- Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system—rationale and benefits. *New England Journal of Medicine*, 366(11), 1051–1056.
- National Commission on Certification of Physician Assistants. (n.d.). *Maintaining certification*. Retrieved October 1, 2023, from <https://www.nccpa.net/maintain-certification/>
- National Conference of Bar Examiners. (2020, November). *Phase 3 report of the Testing Task Force*. National Council of Bar Examiners. <https://nextgenbarexam.ncbex.org/reports/phase-3-report/>
- National Council of Architectural Registration Boards. (2023a, October). *ARE 5.0 format*. National Council of Architectural Registration Boards. <https://www.ncarb.org/pass-the-are/prepare/are-5-0-format>
- National Council of Architectural Registration Boards. (2023b). *Pass the ARE*. <https://www.ncarb.org/pass-the-are>
- National Council of Bar Examiners v. Multistate Legal Studies, 458 F. Supp. 2d 252 (E.D. Pa. 2006).
- National Society of Professional Engineers. (2023). *NSPE: Who we are and what we do*. <https://www.nspe.org/membership/nspe-who-we-are-and-what-we-do>
- Nelson, L. S., Duhigg, L. M., Arnold, G. K., Lipner, R. S., Harvey, A. L., & Reisdorff, E. J. (2019). The association between maintaining American Board of Emergency Medicine certification and state medical board disciplinary actions. *Journal of Emergency Medicine*, 57(6), 772–779.
- Norcini, J. J., Lipner, R. S., & Grosso, L. J. (2013). Assessment in the context of licensure and certification. *Teaching and Learning in Medicine*, 25(Suppl. 1), S62–S67.
- Norcini, J. J., Lipner, R. S., & Kimball, H. R. (2002). Certifying examination performance and patient outcomes following acute myocardial infarction. *Medical Education*, 36(9), 853–859.
- Norcini, J. J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education*, 10(1), 39–59.
- North Carolina Board of Dental Examiners s. FTC 135 S. Ct 1101 (2015).
- O'Neill, T. R., Wang, T., & Newton, W. P. (2022). The American Board of Family Medicine's 8 years of experience with differential item functioning. *Journal of the American Board of Family Medicine*, 35(1) 18–25. <https://doi.org/10.3122/jabfm.2022.01.210208>
- Palmer College of Chiropractic s. Davenport Civil Rights Commission, No. 12-0924 (2014). <https://caselaw.findlaw.com/court/ia-supreme-court/1671310.html>
- Papadakis, M. A., Arnold, G. K., Blank, L. L., Holmboe, E. S., & Lipner, R. S. (2008). Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. *Annals of Internal Medicine*, 148(11), 869–876.

- Parshall, C. G., & Harmes, J. C. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review*, 19(2), 18–25.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Issues in innovative item types. *Practical considerations in computer-based testing* (pp. 70–91). https://doi.org/10.1007/978-1-4613-0083-0_5
- Peabody, M. R., Young, A., Peterson, L. E., O'Neill, T. R., Pei, X., Arnhart, K., Chaudhry, H. J., & Puffer, J. C. (2019). The relationship between board certification and disciplinary actions against board-eligible family physicians. *Academic Medicine*, 94(6), 847–852.
- Phillips, S. E. (2012). Legal issues for standard setting in K–12 educational contexts. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 535–569). Routledge.
- Phillips, S. E. (2017). Legal issue for credentialing examination programs. In S. Davis-Becker & C. Buckendahl (Eds.), *Testing in the occupations: Credentialing policies and practice* (pp. 228–276). Routledge.
- Phillips, S. E., & Camara, W. J. (2006). Legal and ethical issues. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 733–755). Praeger.
- Plake, B. A., & Cizek, C. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 485–502). Routledge.
- Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing interjudge consistency during standard setting. *Educational Measurement: Issues and Practice*, 10(2), 15–16, 22, 25–26.
- Poniatowski, P. A., Dugosh, J. W., Baranowski, R. A., Arnold, G., Lipner, R. S., Dec, G. W., Jr., & Green, M. M. (2019). Incorporating physician input into a maintenance of certification examination: A content validity tool. *Academic Medicine*, 94(9), 1369–1375.
- Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 6, 1–9.
- Professional Certification Coalition. (2023, October). *About PCC*. <https://www.profcertcoalition.org/about-pcc>
- Professional Testing Corporation. (2019). *Standard setting: How to build a panel of subject matter experts*. <https://ptcny.com/standard-setting-build-panel-subject-matter-experts/>
- Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, 23(3), 266–285.
- Ray, M. E., Daugherty, K. K., Lebovitz, L., Rudolph, M. J., Shuford, V. P., & DiVall, M. V. (2018). Best practices on examination construction, administration, and feedback. *American Journal of Pharmaceutical Education*, 82(10), 7066. <https://doi.org/10.5688/ajpe7066>

- Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14(4), 369–415.
- Raymond, M. R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues and Practice*, 21(3), 25–37.
- Raymond, M. R. (2015). Job analysis, practice analysis, and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 144–164). Routledge.
- Raymond, M. R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 181–223). Lawrence Erlbaum Associates.
- Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts methods, and perspectives* (pp. 119–157). Lawrence Erlbaum Associates.
- Romboy, D. (2012, August 10). *Court sides with woman in hair braiding case*. KSL Broadcasting. <https://www.ksl.com/article/21638994/court-sides-with-woman-in-hair-braiding-case>
- Rosenfeld, M., Thornton, R., & Shimberg, B. (1983). Job analysis of licensed psychologists. *Professional Practice of Psychology*, 4(2), 17–24.
- Rudner, L. M., & Gao, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology*, 12, 1–12.
- Schmitt, K. (1995). What is licensure? In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 3–32). Buros Institute of Mental Measurements. <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1004&context=buroslicensure>
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), 38–48.
- Section 504 of the Rehabilitation Act [Section 504], 29 U.S.C. § 701 *et seq.* (1973).
- Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4(4), 447–467.
- Shimberg, B. (1982). *Occupational licensing: A public perspective*. ETS.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47, 150–174.
- Sinharay, S., Puhon, G., & Haberman, S. J. (2011). An NCME instruction module on subscores. *Educational Measurement: Issues and Practice*, 30(3), 29–40.
- Smiley, W. (2019, April). *An examination of classification accuracy in a continuous testing assessment framework* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, ON, Canada.
- Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, 22(4), 26–33.
- Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement*, 25(4), 259–285.

- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.).
- Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 13, 265–276.
- Swanson, D. B., & Hawkins R. E. (2017). Using written exams to assess medical knowledge and its application. In E. S. Holmboe, S. J. Durning, & R. E. Hawkins (Eds.), *Practical guide to the evaluation of clinical competence* (2nd ed., pp. 42–59). Mosby Elsevier.
- Taylor, K. (2017, March 13). Regents drop teacher literacy test seen as discriminatory. *New York Times*. https://www.nytimes.com/2017/03/13/nyregion/ny-regents-teacher-exams-alst.html?mwrs&_r=0
- Tamblyn, R., Abrahamowicz, M., Dauphinee, W. D., Hanley, J. A., Norcini, J., Girard, N., Grand'Maison, P., & Brailovsky, C. (2002). Association between licensure examination scores and practice in primary care. *Journal of the American Medical Association*, 288(23), 3019–3026.
- Teirstein, P. S. (2015). Boarded to death—Why maintenance of certification is bad for doctors and patients. *New England Journal of Medicine*, 372(2), 106–108.
- Teirstein, P. S., & Topol, E. J. (2015). The role of maintenance of certification programs in governance and professionalism. *Journal of the American Medical Association*, 313(18), 1809–1810.
- Thompson, N. (2018, May 23). *What is the Hofstee method for setting cutscores?* Assessment Systems Corporation. <https://assess.com/what-is-the-Hofstee-method-for-setting-cutscores/>
- Title VII of the Civil Rights Act, 42 U.S.C. § 2000e *et seq.* (1964).
- Uijtdehaage, S., & Schuwirth, L. W. T. (2018). Assuring the quality of programmatic assessment: Moving beyond psychometrics. *Perspectives on Medical Education*, 7(6), 350–351.
- Ura, A. (2019, May 27). Texas plumbing board and laws abolished after legislative strife. *Texas Tribune*. <https://www.texastribune.org/2019/05/27/texas-plumbing-board-and-laws-abolished-after-legislative-strife/>
- Vandergrift, J. L., Gray, B. M., Barnhart, B. J., Lynn, L. A., & Lipner, R. S. (2020). Opportunities for maintenance of certification to better reflect scope of practice among medical oncologists. *JCO Oncology Practice*, 16(8), 641–648.
- Vandergrift, J. L., Weng, W., & Gray, B. M. (2021). The association between physician knowledge and inappropriate medications for older populations. *Journal of the American Geriatric Society*, 69(12), 3584–3594.
- van der Linden, W. J., & Li, J. (2016). Comment on three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement*, 40(8), 641–649.
- von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, 54(1), 3–11.
- Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K., & Thissen, D. (2001). Augmented scores—“borrowing strength” to compute scores

- based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Routledge.
- Waldren, S. E., Agresta, T., & Wilkes, T. (2017). Technology tools and trends for better patient care: beyond the EHR. *Family Practice Management*, 24(5), 28–32.
- Way, W. D., Steffen, M., & Anderson, G. A. (2014). Developing, maintaining, and renewing the item inventory to support CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (3rd ed., pp. 143–164). Routledge.
- Way, W. D., & Gialluca, K. A. (2017). Interpreting the meaning of test scores. In C. Buckendahl & S. Davis-Becker (Eds.), *Testing in the professions* (pp. 105–122). Taylor & Francis.
- Wendt, A., & Harmes, J. C. (2009). Evaluating innovative items for the NCLEX, Part 1: Usability and pilot testing. *Nurse Educator*, 34(2), 56–59.
- White, W. D. (2014). Professional self-regulation in medicine. *Virtual Mentor*, 16(4), 275–278.
- Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Psychology Press.
- Zaglaniczny, K. L. (1993). Council on certification professional practice analysis. *Journal of the American Academy of Nurse Anesthetists*, 61(3), 241–255.
- Zapata-Rivera, D. (Ed.). (2018). *Score reporting research and applications*. Routledge.
- Zenisky, A. L., & Hambleton, R. K. (2015). A model and good practices for score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 585–602). Routledge.
- Zhou, Y., Sun, H., Culley, D. J., Young, A., Harman, A. E., & Warner, D. O. (2017). Effectiveness of written and oral specialty certification examinations to predict actions against the medical licenses of anesthesiologists. *Anesthesiology*, 126(6), 1171–1179.
- Zhou, Y., Sun, H., Macario, A., Keegan, M. T., Patterson, A. J., Minhaj, M. M., Wang, T., Harman, A. E., & Warner, D. O. (2018). Association between performance in a maintenance of certification program and disciplinary actions against the medical licenses of anesthesiologists. *Anesthesiology*, 129(4), 812–820.
- Zhou, Y., Sun, H., Macario, A., Keegan, M. T., Patterson, A. J., Minhaj, M. M., & Warner, D. O. (2019). Association between participation and performance in MOCA Minute and actions against the medical licenses of anesthesiologists. *Anesthesia & Analgesia*, 129(5), 1401–1407.

NOTES

1. Teacher certification represents an interesting example of ambiguity between licensure and certification. States require certification for individuals teaching in public schools, but there is often no similar restriction for teaching outside the public system. This is similar to physician certification in the United Kingdom, where

physicians can practice without being a member of one of the Royal Colleges but cannot receive reimbursement from the National Health Service.

2. As with many issues related to licensure and certification, there are exceptions—or at least gray areas. For example, the USMLE comprises three steps. Successful completion of Steps 1 and 2 typically is required to enter residency training, and in some states individuals receive a formal training license. Because of this special use, these first two steps focus on what might be viewed as entry-level skills. A full unrestricted license is, however, only awarded after completing the third step in the examination process—the focus of that step is not constrained in the same way.
3. Work samples are another common item type in personnel assessments for certification.
4. In addition, it is important that decisions about feedback be made at the test development phase (i.e., as an intentional part of the overall testing process) rather than as an additional consideration after the test has been developed and/or administered.

Assessment of Social-Emotional, Soft, Character, Behavioral, and Intrapersonal and Interpersonal Skills

Patrick C. Kyllonen
ETS

Jiyun Zu
ETS

Intelligence plus character—that is the goal of true education.

Martin Luther King Jr.

We have long recognized that schools develop character along with mathematics and language skills. Students are taught, at least implicitly, to pay attention to the instructor, to put effort into their work, to show up and complete assignments on time, to be considerate of other students, to avoid arguing with the teacher and fighting with other students, to be honest and truthful, and to develop a love of learning. As a society, we believe that such behaviors and attitudes are important for the future success of students—success in education, in the workforce, and in life, generally. However, these beliefs have not typically translated into policy or practice in educational testing.

Testing reflects policy priorities and influences what is taught and learned (Frederiksen, 1984; Levin, 2012). Since 2001, federal and state policies have centered around evaluation of students' knowledge of subject areas, as seen in the Every Student Succeeds Act and its predecessor, the No Child Left Behind Act of 2001, as well as in our *Nation's Report Card* (National Assessment of Educational Progress). International assessments such as the Program for International Assessment (PISA), which began in 2000; Trends in International Math and Science Study (TIMSS), beginning in 1995; and Progress in International Reading Literacy Study (PIRLS), beginning in 2001, along with national accountability assessments in most countries of the Organisation for Economic Co-operation and Development (OECD) all focus on curricular content knowledge. We evaluate teachers with value-added models based entirely on student gains in academic achievement, as measured by curricular tests (McCaffrey et al., 2003).

Only recently has there been an explicit attempt to move beyond a general acknowledgment of the school's role in building character to embracing character development as part of the core school mission warranting its monitoring. In the United States, a subset of 10 California school districts (the California Office to Reform Education) received a waiver from the federal government to implement their own accountability assessment, which included social-emotional learning skills (West, 2016). The District of Columbia Public Schools is administering social and emotional learning (SEL) survey measures to assess readiness (District of Columbia Public Schools, 2017). Many states have by now published social-emotional learning standards paralleling mathematics, English language arts, and science standards (Collaborative for Academic, Social, and Emotional Learning [CASEL], 2020a). Chile's ministry of education is mandated to include social-emotional skills as part of its national accountability assessment (Ministerio de Educacio, 2016); other countries are following suit (Bravo-Senzana et al., 2023). The OECD recently launched its Study on Social and Emotional Skills parallel to PISA, but designed to monitor progress in student development of social-emotional skills, rather than cognitive skills (OECD, 2019a). Many organizations have issued position papers on the importance of SEL (Aspen Institute, 2018; Atwell & Bridgeland, 2019; OECD, 2015; Salzburg Global Seminar, 2018; Schanzenbach et al., 2016). Online resources are now available to find and evaluate

social-emotional learning assessments, including Rand's (Hamilton et al., 2018) Education Assessment Finder, CASEL's (2020b) SEL Assessment Guide, California's Guide to Social and Emotional Learning Resources (California Department of Education, 2018), and others. What has changed?

EVIDENCE FOR THE IMPORTANCE OF INTERPERSONAL AND INTRAPERSONAL SKILLS

Benefits-of-Education Studies

One important source for change was economists working from a human capital theory framework. Human capital theory (Becker, 1994) is the idea that individuals are rewarded for their productivity (e.g., measured as earnings, or labor market outcomes more generally, including employment and nonincarceration), which is driven by the set of skills, knowledge, and health individuals bring to the labor market. These attributes are affected by education, on-the-job training, medical care, and other factors. This framework enables quantifying the value of investments in education on economic productivity, the rate of return for a high school or college education (Card, 1999). A finding is that educational attainment is an important causal determinant of employment and earnings (private returns to education; Barrow & Rouse, 2005; Card, 1999) as well as national economic strength (social returns to education; Moretti, 2005), hence the public investment in education. However, the benefit of education on labor market outcomes is only partly due to the gains in cognitive skills associated with education, perhaps only about 20% (Bowles et al., 2001). Most of the benefit is due to factors other than the cognitive skills acquired, as measured with standardized tests, hence the attribution to noncognitive skills. Recent evidence shows that returns to cognitive ability have even gone down since the 1980s, with a 30% to 50% larger effect of wages in the 1980s compared to the 2000s (Castex & Dechter, 2014).

Another key demonstration of the benefits of education not accounted for by cognitive ability was Heckman and Rubenstein's (2001) finding that General Educational Development (GED) holders, individuals who failed to complete high school but scored comparably to high school degree holders on cognitive tests and therefore were similar in cognitive skills, experienced very different labor market outcomes. GED holders were far more likely to be unemployed and have trouble with the law (Heckman & Rubinstein, 2001; Heckman et al., 2014). The signaling value of a high school diploma is a potential alternative explanation for these findings—employers value degrees—but empirical studies support a human capital interpretation over a signaling interpretation (Clark & Martorell, 2014).

A third source of findings concerns the value of early childhood education. Although such programs were motivated with the goal of boosting cognitive ability for children living in impoverished environments, a common finding was that cognitive ability gains tended to fade out after a few years, as determined by comparison with control groups (Bailey et al., 2017). However, long-term treatment effects have consistently

been realized in many positive outcomes, including increased educational attainment and employment and less incarceration (Barnett, 1996).

The significance of all three of these study clusters is that they demonstrate the importance of noncognitive skills—skills associated with education that were not related to the cognitive gains education produces. However, aside from the cognitive versus noncognitive distinction, these studies tended not to identify the specific nature of the noncognitive factors associated with the education effect. A question is, What are these “noncognitive” benefits of education?—What is it that teachers teach and students learn, other than mathematics, reading, science, and other subject matters, that are of value in school, to employers, and in life?

Predictions-From-Measures Studies

Studies have been conducted by economists and psychologists examining the predictions of educational, workplace, and life outcomes from self- or other-report surveys. In higher education, a meta-analysis of psychosocial and study skill factors by Robbins et al. (2004) found the highest correlates of grades to be academic self-efficacy and achievement motivation and those of retention to be academic goals, academic self-efficacy, and academic-related skills. These factors were predictive after controlling for socioeconomic status, cognitive ability (test scores), and high school grades. Richardson et al. (2012) similarly found academic self-efficacy to be among the highest correlates of grades, along with effort regulation, need for cognition, and grade goal. In economics, Lindqvist and Vestman (2011) found that a composite measure based on ratings by a clinical psychologist of 18-year-old males predicted low earnings and chronic unemployment 20 years later. Segal (2013) found that misbehavior ratings of eighth-grade students by their teachers predicted earnings 20 years later after controlling for educational attainment and cognitive test scores.

These studies are suggestive of the importance of noncognitive factors but were also limited by the lack of a reliable strategy for categorizing noncognitive predictors. Beginning in the late 1980s and into the 1990s, the Big Five or five-factor model of personality arose in prominence (P. T. Costa & McCrae, 1992; Digman, 1990; Goldberg, 1993; John & Srivastava, 1999). We discuss this model further in the section “What Are the Key Skills?” but for our purposes here, it suffices to say that the model posits that individual personality descriptions are captured by five independent dimensions, Extroversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. As in many realms, particularly measurement, standardization leads to progress (Cochrane, 1966), and the adoption of the five-factor model had a major impact on the value and frequency of prediction studies. Meta-analyses of workplace outcomes (Barrick & Mount, 1991; Dudley et al., 2006; Rothmann & Coetzer, 2003; Salgado, 1997; Tett et al., 1991), lifetime earnings outcomes (Gensowski, 2018), general life outcomes (B. W. Roberts et al., 2007), and educational outcomes (Noffle & Robins, 2007; Poropat, 2009; Salgado & Tauriz, 2014) have been conducted within the Big Five framework, leading to findings particularly of the importance of Conscientiousness (i.e.,

being organized, responsible, and achievement striving) on school, life, and workforce outcomes (Almlund et al., 2011; Levin, 2012; National Research Council, 2012; Ozer & Benet-Martinez, 2006).

Note on the Relative Importance of Cognitive Versus Noncognitive Skills

Cognitive ability is sometimes treated as a control variable to examine the incremental predictiveness of noncognitive variables (Nofle & Robins, 2007; Robbins et al., 2004; Segal, 2013). However, several studies have also compared the prediction given by cognitive versus noncognitive variables. Poropat (2009) and Richardson et al. (2012) found that some of the noncognitive factors have as strong a relationship with outcomes as cognitive abilities do. Poropat (2009) computed correlations of academic performance with intelligence, $\rho = .25$ (47 studies, all levels), and Conscientiousness, $\rho = .22$ (138 studies, all levels), findings in line with findings from the workforce (Levin, 1989; Salgado & Tauriz, 2014). Borghans et al. (2016) examined four large longitudinal data sets to show that personality, as expressed in grades and achievement scores, was more predictive of various life outcomes (e.g., wages, education, arrests, life satisfaction, civic participation) than IQ scores were. B. W. Roberts et al. (2007) found that the correlations between personality measures and a variety of life outcomes (mortality, divorce, occupational attainment) were of the same magnitude as the correlations between cognitive ability measures and socioeconomic status measures with those outcomes.

One complication in reporting simple correlations is correcting for range restriction—Levin (2012) pointed out that such corrections can distort findings. Selection on cognitive and noncognitive ability can operate differently—selecting on cognitive ability is often direct and affects the variance of cognitive test scores on the selected sample. Selecting on noncognitive factors is often indirect and often unmeasured. However, a persistent finding is that within a cohort, the correlation between noncognitive scores and outcomes is stronger for those in the lower half of the ability distribution (Lindqvist & Vestman, 2011; Segal, 2013). Noncognitive ability appears to be particularly important for students who leave school early and have difficulties maintaining employment.

A challenge in comparing the relative importance of cognitive and noncognitive skills is that measures of them call on both skills so their effects are difficult to disentangle (Borghans et al., 2011, 2016; Kyllonen & Kell, 2018). Being unmotivated temporarily or characteristically can produce low cognitive test scores and comprehension difficulties can affect responses on a personality survey.

Workplace-Demands Studies

Employer surveys ask employers what skills they look for in new graduates or how important various skills were for job performance. The findings from such surveys have been consistent. Casner-Lotto and Barrington (2006) found that what they called *applied skills* (e.g., critical thinking/problem-solving, oral and written communications, teamwork/collaboration, diversity, leadership, professionalism/work ethic) were

far more likely to be considered “very important” to success at work by employers compared to *basic knowledge/skills* (e.g., English language, reading comprehension, mathematics, science), particularly for high school and 2-year graduates. For these two groups, the two skills most likely to be rated as most important were professionalism/work ethic and teamwork/collaboration. Other surveys have found similar results, both domestically (Cengage, 2019; National Association of Colleges and Employers, 2018) and internationally (Barton et al., 2012).

A question is whether employer surveys provide a distorted picture of the skills sought. Employers might say *X* is important when asked in a survey, while looking for *Y* during recruiting. Rios et al. (2020) examined 142,000 job advertisements and found that 70% of job ads requested at least one of what they called “21st-century skills,” with the top skills being oral and written communication, collaboration, problem-solving, communication skills, social intelligence, and self-direction. These results are largely in line with results from the employer surveys, although terminology differences make them somewhat hard to compare. Employers routinely assess these skills using personality tests, structured interviews, situational judgment tests, game-based assessments, and reference checking (U.S. Office of Personnel Management, 2018b).

Another kind of workplace demand study has involved examining the actual activities workers engage in on the job. These studies typically have the theme of effects of technology on the nature of work. An influential early study by Autor et al. (2003) examined jobs (using the Dictionary of Occupational Titles database) in the years 1960 to 1998 and found that the effect of technology was to substitute for routine cognitive and manual tasks but complement activities involving nonroutine analytical and interpersonal tasks. This has led to a decline in manual tasks (routine and nonroutine) and a rise in nonroutine analytical and interpersonal tasks, thereby increasing the demand for the latter. The analysis was later extended to 2009, showing even clearer trends, but also showing a plateauing of nonroutine tasks (Autor & Price, 2013). Deming (2017) conducted a similar study covering the years 1980 to 2012 and showed that jobs requiring social interactions have grown considerably, showing a trend for an increasing “social skills” premium. Weinberger (2014) presented similar findings using Dictionary of Occupational Titles data showing a growth in employment and earnings for occupations requiring high levels of both cognitive and social skills compared to those requiring only one or the other. Burrus et al. (2013) applied a principal component analysis to O*NET ratings, identifying 15 competency components (mixtures of cognitive and noncognitive competencies), and showed that the most important competencies by both importance ratings and correlations with earnings were what they labeled problem-solving, fluid intelligence, teamwork, achievement/innovation, and communication skills.

There also have been studies attempting to project skill demands for the workplace of the future. A workshop conducted by the National Research Council (2008) reinforced the findings of the growing importance of broad interpersonal and social skills, noting the same conclusion in the earlier SCANS Commission report (Kane

et al., 1990) and suggesting the addition of adaptability. A study by Frey and Osborne (2017) attempted to predict the probability of computerization for 702 occupations in the United States, based on O*NET data, concluding that close to half of U.S. employment was at risk. They showed that the most resistant occupations were ones involving perception (finger and manual dexterity), creativity (originality and fine arts), and social intelligence (social perceptiveness, negotiation, persuasion, and caring for others), particularly when in high-wage jobs requiring high levels of educational attainment. A study by McKinsey Global Institute (Manyika et al. (2017) concluded that future occupations will require more interactions and management, as well as more social-emotional, creative, and logical reasoning abilities. A recent edited volume by Oswald et al. (2019) explored educational implications for the future of work (Hilton, 2019).

These three clusters of studies—benefits of education, predictions from measures, and workplace demands—all point to a growing recognition of the importance of non-cognitive, character, or social-emotional skills. There also appears to be evidence that the importance of these skills is likely to increase in the future as a result of changes in the workforce related to technology. This highlights the importance of the topic for educational measurement.

A Note on Terminology

In the title of this chapter, we refer to interpersonal and intrapersonal skills and other names; in the chapter, we use various terms to refer to the same general set of attributes. Terminology in this field is a source of frustration and confusion (Duckworth & Yeager, 2015; Reeves & Venator, 2014). SEL or social-emotional competencies or skills is terminology that has taken hold in K–12 education circles (CASEL, 2020b). It perhaps supersedes character or character skills, although character education is still a prominent reference. For a long time, What Works Clearinghouse maintained a separate Character Education category (Institute of Educational Sciences, 2007), although that appears to now be replaced by a Behavior category (Institute of Educational Sciences, 2020). In higher education, interpersonal and intrapersonal skills are widely used because of the National Academies of Science's (2017) adoption of that terminology. Also used is “hard-to-measure” skills or competencies (Stecher & Hamilton, 2014) and 21st-century competencies (Soland et al., 2013). In the workforce, soft skills (Kyllonen, 2013) and behavioral competencies (Society for Human Resource Management, 2014) are widely used, as is emotional intelligence (Goleman, 1995) (Emotional intelligence also has a more technical meaning; J. D. Meyer et al., 2008.) Economists have historically referred to noncognitive skills, to contrast with the cognitive skills measured with IQ and standardized achievement tests (although Messick, 1978, used the term noncognitive, noting its problems, but blaming its necessity on Bloom's 1956 usurpation of cognitive to define a particular set of skills). Industrial-organizational psychology refers to “other factors” as seen in the acronym KSAO, for knowledge, skills, abilities, and other factors.

Skills are prominent in economic or human capital theory discussions as a shorthand for whatever is learned in school or in training that has value in the workforce. But measurement psychologists have historically distinguished knowledge, skills, and abilities and added attitudes, values, beliefs, and behaviors, as seen in *Buros Mental Measurements Yearbook* (Carlson et al., 2020). Twenty-first-century skills, skills for the new economy, and others appear in the literature to reflect the changing demands of school and the workplace as a result of changes in technology (Oswald et al., 2019). *Skill* itself as a term may be thought of as controversial to describe some attributes, but its evolving widespread acceptance is shown in a *New York Times* opinion piece by U.S. senator Jeff Flake (2017), titled “We Need Immigrants With Skills, But Working Hard Is a Skill.” Personality is also used. However, the concept of personality, like the concept of intelligence, is associated with heredity and permanence, and so educators have tended to shy away from this terminology, believing personality to be outside the control of teachers and the educational system. As discussed in the next section, the foundation for this belief may be chipping away.

MALLEABILITY OF PERSONALITY

A recurring problem in the field of the assessment of noncognitive skills in education is the widespread belief that personality or character traits are fixed, and therefore outside the realm of education. Personality traits have been defined as the “relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances” (B. W. Roberts, 2009). Numerous studies have shown that personality traits are rank-order stable over decades (Ferguson, 2010) and heritable (Vukasovic & Bratko, 2015). However, this is true of achievement as well (Rimfeld et al., 2018), and also of social cognitive constructs, such as interests, self-concept, and academic effort, popularly believed to be more malleable, but which have been found to be no more so than the Big Five personality traits (Rieger et al., 2017).

Damian et al. (2019) tracked personality change in 1,795 adolescents tested 50 years later, finding moderate stability of personality across that period ($r = .31$ corrected for measurement error). But they also found growth of about half a standard deviation for personality factors over the period, in the direction of maturity (over time, greater Emotional Stability, Conscientiousness, Agreeableness, Openness).¹ Specific activities, such as psychotherapy, can have relatively swift and substantial effects on personality change (B. W. Roberts et al. 2017). Social-emotional learning interventions in school have similar and lasting effects on students’ personality as well as academic and other outcomes (Durlak et al., 2010, 2011; Taylor et al., 2017). Together, these findings challenge beliefs about the permanence of noncognitive factors or their inaccessibility to education, whether they are called personality traits, social-emotional competencies, noncognitive skills, or something else. Like habits, personality can change; personality is relatively enduring but not fixed.

WHAT ARE THE KEY SKILLS?

Considerable evidence supports the importance of noncognitive skills in school, the workplace, and life. What are those skills? What noncognitive skills are most important, and what skills are the ones driving educational and workforce outcomes? What skills should be given attention in education, and what skills should be monitored for growth and development?

Two specific skills have received an enormous amount of attention in education circles. One is grit (Duckworth, 2016), and the other is growth mindset (Dweck, 2006). Since 2016, it seems to have been almost impossible to attend a meeting of educational practitioners or policy makers without the word *grit* being mentioned. Grit seems to play an outsized role, serving as a stand-in for all the skills that are missed by standardized cognitive tests. Growth mindset has played a similar role and has received a comparable number of mentions in practitioner and policy discussions. A solid scientific literature attests to the importance of the constructs of grit and growth mindset on outcomes. However, something else may be at play. Both constructs have simple but compelling narratives—for grit, work hard in a sustained way and you will succeed; for growth mindset, believe that your successes and failures are a result of the effort put in, rather than your intelligence or external factors, and by doing so you will respond more productively to successes and failures (“the power of yet”). The two constructs also have charismatic advocates with popular TED talks (Duckworth, 2013, 50 languages, over 19 million views; Dweck, 2014, 42 languages, over 10 million views) and best-selling books (Duckworth, 2016, 21 weeks on the *New York Times* bestseller list; Dweck, 2017).

There have been meta-analyses of the constructs. Credé et al. (2017) examined 584 effect sizes from 88 samples and 66,807 individuals and showed that Duckworth’s grit was uncorrelated with cognitive ability ($\rho = .05$)² and highly correlated with Conscientiousness from the Big Five ($\rho = .84$). Grit predicted grades, both in high school ($\rho = .16$) and in college ($\rho = .17$; the study did not report grit’s incremental prediction beyond general cognitive ability, but given the low correlation between grit and cognitive ability, the result likely would have been similar). Grit could be separated into two correlated constructs, perseverance and consistency (correlated $\rho = .66$), and perseverance was shown to be the stronger correlate of grades ($\rho = .26$ and $\rho = .10$, respectively). These findings are largely in line with Poropat’s (2009) findings and others reviewed in the previous section.

Several meta-analyses of the growth mindset work have been conducted. Sisk et al. (2018) examined $k = 273$ studies with over $N = 365,000$ students and showed that the correlation between growth mindset and achievement was $\rho = .12$, and only about 37% of the studies found a correlation greater than zero. They also found that this relationship was not moderated by academic risk, socioeconomic status, or test score versus grades outcomes. But it was moderated by age, such that the correlation was zero for adults. Sisk et al. (2018) also examined the effect size of mindset interventions, finding $k = 43$ studies ($N = 57,155$ participants) attempting to improve growth mindset. Of the 43 studies, 5 of them showed significant improvement, with an estimated effect size of

$d = .08$. Age, at-risk status, intervention type, and intervention length did not moderate effect size. But there were indications that lower socioeconomic groups benefited more and that reading-based interventions were more effective than computer programs. Yeager et al.'s (2019) randomized control trial ($N = 6,320$ relatively low-achieving students) found a similar estimate, $d = .11$, on grades. Lazowski and Hulleman (2016; $k = 92$, $N = 38,377$) estimated a much higher $d = .49$, although they included both randomized and quasi-experimental designs and a larger group of "motivation interventions." Other than the latter estimate, these effect sizes are "small" by Cohen's (1992) rules of thumb, although Yeager et al. (2019) argued that $d = .2$ should be considered a large effect and that their intervention attained a substantial proportion of that effect.

However, as can be seen from Hattie's (2009) review of effect sizes related to student achievement, there may be many more potentially important factors than grit and growth mindset. Of Hattie's 252 effect sizes, 225 of these showed $|d| > .11$ (Yeager et al.'s 2019 estimate is based on a randomized control trial, whereas Hattie's also included observational data).

At the time of this writing, the CASEL SEL Assessment Guide (2020b) listed 157 constructs from 26 assessments. An organizing framework is clearly needed. John and De Fruyt (2015) provided a way to organize all "21st-century constructs" identified in a review (Trilling & Fadel, 2009). Table 19.1 provides the results of a principal component analysis of self-ratings on these constructs by 350 University of California undergraduates. Five components were extracted that align with the Big Five, suggesting that the five-factor model is a robust characterization of human self-descriptions, regardless of the origins of those self-descriptions.

MOST IMPORTANT SKILLS TO ASSESS: FRAMEWORKS

Various proposals suggest how best to organize the large pool of noncognitive constructs and to highlight the most important ones. We review those that have the most impact on education discussions.

Collaborative for Academic, Social, and Emotional Learning (CASEL): Five-Dimension Framework

CASEL has influenced policy and practice related to social and emotional skills in schools in the United States, particularly in pre-K–12. The CASEL website states that "CASEL was formed in 1994 with the goal of establishing high-quality, evidence-based social and emotional learning (SEL) as an essential part of preschool through high school education." CASEL proposes (a) to conduct research on the efficacy of SEL, (b) to assist in practice and implementation, and (c) to promote legislation at the local, state, and federal levels.

CASEL worked with the Illinois State Board of Education to implement SEL student learning standards for three broad goals of (a) developing self-awareness and

Table 19.1 Principal Component Analysis of 21st-Century Skills and Big Five Alignment

Factor 1: Collaboration (A)	Factor 2: Task Performance (C)	Factor 3: Emotion Regulation (N)	Factor 4: Engagement With Others (E)	Factor 5: Open- mindedness (O)
Compassion	Self-discipline	Self-confidence	Social connection	Curiosity
Care	Focus	Self-esteem	Teamwork	Inquisitiveness
Cooperation	Perseverance	Decisiveness	Social awareness	Willingness to try new ideas
Kindness	Self-control, school	Tackling tough problems	Public speaking	Receptivity
Respect for others	Grit	Cheerfulness	Assertiveness	Innovation
Empathy	Organization	Happiness	Leadership	Vision
Tolerance	Diligence	Optimism	Courage	Insight
Fairness	Precision	Tranquility	Charisma	Tinkering (inventing)
Trust	Dependability	Balance	Speaking out/taking a stand	Learning from mistakes
Forgiveness	Reliability	Stability	Bravery	Excitement creating something new
Gratitude	Consistency	Equanimity	Enthusiasm	Appreciating beauty in the world
Appreciation of others	Trustworthiness	Self-compassion	Passion	Living in harmony with nature
Living in harmony with others	Goal orientation	Self-kindness	Zeal	Spirituality
Interconnectedness	Motivation		Inspiration	Mindfulness
Inclusiveness	Work ethic		Spunk	Existentiality
	Effort		Spontaneity	Awe
	Productivity		Playfulness	Wonder
			Humor	Reverence
				Self-reflection
				Self-awareness
				Consciousness
				Self-actualization
				Authenticity

Note. Column headings indicate aligning Big Five factor: A = Agreeableness, C = Conscientiousness, N = Neuroticism, E = Extraversion, O = Openness. Adapted from “Education and Social Progress: Framework for the Longitudinal Study of Social and Emotional Skills in Cities” (EDU-CERI-CD(2015)13.en), by O. P. John & F. De Fruyt, 2015. OECD Publishing.

self-management skills, (b) using social awareness and interpersonal skills to establish and maintain positive relationships, and (c) demonstrating decision-making skills and responsible behaviors in personal, school, and community contexts (Durlak et al., 2011). The components of these goals—self-awareness, self-management, social awareness, relationship skills, and responsible decision-making—have become the major constructs around which CASEL’s work is organized. The “CASEL 5” did not result from a scientific analysis of either SEL programs or individual differences, but instead informally emerged to characterize aspects and points of emphasis of various SEL programs, using language that policy makers found useful. SEL language has been adopted widely in government circles, such as in state and national standards, television programs, and other media.

CASEL maintains an extensive website with resources and links. CASEL focuses on SEL programs rather than assessments, but hosted an assessment work group for several years, resulting in several useful reports (Assessment Work Group, 2019) and resources, including the SEL Assessment Guide (CASEL, 2020b), along with a blog, a comparison of frameworks (Blyth et al., 2019), an alignment tool, webinars, and assessments emerging from several years of design challenges. CASEL has produced summaries and meta-analyses on the efficacy of SEL programs (Durlak et al., 2010, 2011; Mahoney et al., 2018; Taylor et al., 2017), a handbook on research and practice (Durlak et al., 2015); and position papers based on research findings and working with stakeholders on the CASEL website.

National Academy of Sciences/National Research Council

The National Research Council (2008, 2011, 2012) and its renamed successor, the National Academies of Sciences, Engineering, and Medicine (2017), conducted a series of workshops and consensus reports on interpersonal and intrapersonal skills. The 2012 consensus report reviewed several 21st-century skills frameworks and created a synthesized framework, organized around models from individual differences studies in abilities (Carroll, 1993) and personality (John & Srivastava, 1999). The taxonomy posited a set of cognitive (cognitive processes and strategies, knowledge, creativity), interpersonal (teamwork and collaboration, leadership), and intrapersonal (intellectual openness, work ethic, positive core self-evaluation) competency clusters. Each competency cluster (e.g., teamwork and collaboration) was in turn cross-walked to 21st-century terminology taken from the source documents (communication, collaboration, teamwork, cooperation, coordination, interpersonal skills, empathy, perspective taking, trust, service orientation, conflict resolution, negotiation), an O*NET descriptor or skill (Peterson et al., 1999; social skills), and a basic factor from the individual differences literature (e.g., Agreeableness). The value of organizing around existing models of abilities and personality is that such models provide an empirical foundation for categorizing constructs, in much the same way that John and De Fruyt’s (2015) analysis does (see Table 19.1). This is a way to address the problem of jingle (a single term to describe multiple constructs) and jangle (multiple terms to describe the same

construct; Reeves & Vanator, 2014). It also provides a way to categorize interventions, with an expectation that specific construct interventions will provide spillover effects on within-category constructs in accord with a transfer gradient.

The National Academies (2017, p. 2) also convened a group to

examine how to assess interpersonal and intrapersonal competencies (e.g., teamwork, communication skills, academic mindset, and grit) of undergraduate students for different purposes . . . to include identifying a range of competencies that may be related to postsecondary persistence and success, and that evidence indicates can be enhanced through intervention. (p. 2)

They identified eight competencies: (a) “behaviors related to conscientiousness . . . to self-control, hard work, persistence, and achievement orientation”; (b) “sense of belonging—a student’s sense that he or she belongs at a college, fits in well, and is socially integrated”; (c) “academic self-efficacy—a student’s belief that he or she can succeed in academic tasks”; (d) “growth mindset—a student’s belief that his or her own intelligence is not a fixed entity, but a malleable quality that can grow and improve”; (e) “utility goals and values—personal goals and values that a student perceives to be directly linked to the achievement of a future, desired end”; (f) “intrinsic goals and interest—personal goals that a student experiences as rewarding in and of themselves, linked to strong interest”; (g) “prosocial goals and values—the desire to promote the well-being or development of other people or of domains that transcend the self”; and (h) “positive future self—a positive image or personal narrative constructed by a student to represent what kind of person he or she will be in the future” (pp. 5–6). The report stated that “self-report methods, with their known limitations, predominated in the assessments of the eight competencies” (p. 8) and that “analysis of the quality of the assessments used in the intervention studies revealed spotty attention to reliability and almost no reported evidence of validity or fairness.” (p. 8).

OECD’s SSSES framework

OECD manages the PISA (OECD, 2019b) and the Program for the Assessment of Adult Competencies (PIAAC), among others. PISA and PIAAC are surveys measuring respondents’ cognitive proficiency and include background questionnaires. OECD recently launched the Study on Social and Emotional Skills, an international cross-sectional (ages 10 and 15) survey of students, teachers, and parents focusing on social and emotional skills, conducted in 12 city sites (Kankaraš, 2017; OECD, 2019a). The long-term goal is to monitor the growth of social-emotional skills from school to the labor force in a policy-relevant, feasible, valid, reliable, comparable, ethical, cost-effective, and sustainable way.

The project has published several framework documents (Chernyshenko et al., 2018; John & De Fruyt, 2015; Kankaraš & Suarez-Alvarez, 2019; OECD, 2015). OECD (2015) provided the justification for the study with a comprehensive literature review, concluding that social and emotional skills, as important as they are, were hard to measure, and consequently teachers, schools, and policy makers

do not know “if their efforts at developing these skills are paying off,” and that these skills “are seldom taken into account in school and university admissions decisions” (p. 13).

OECD (2019a) proposed the Big Five as an organizing framework because of positive research on the psychometric qualities of Big Five measures (reliability, predictions of outcomes) and a sound justification for five independent dimensions. Even with only five dimensions, the model is comprehensive, particularly when lower order correlated dimensions are added (Dudley et al., 2006; John & Srivastava, 1999; Paunonen & Ashton, 2001). Also, the Big Five has a scientific underpinning in the lexical hypothesis (Goldberg, 1993), that language evolves to accommodate the most important and salient differences between people in the form of descriptors. There are competitors to the Big Five, such as the six-dimension hypothesis (Lee & Ashton, 2004), but this is a relatively minor variation. Major alternative explanatory models are not readily available. Seeming alternatives, such as 21st-century skills, can easily be shown to be largely captured by the Big Five, as John and De Fruyt’s (2015) analysis showed (see Table 19.1).

The OECD (2019a) framework proposed five domains with several facets within each domain: task performance (achievement motivation, responsibility, persistence, self-control), emotional regulation (stress resistance, optimism, emotional control), collaboration (empathy, trust, cooperation), open-mindedness (tolerance, creativity, curiosity), and engaging with others (sociability, assertiveness, energy). These are the Big Five factors (Conscientiousness, Neuroticism, Agreeableness, Openness/Intellect, and Extraversion, respectively) with more palatable names. The framework adds a sixth dimension, compound skills, which includes critical thinking, metacognition, and self-efficacy. Unlike some of the frameworks reviewed here, this framework is not merely notional or a helpful organizing scheme, but a structural model and, with the assessments being administered, a measurement model, which can be evaluated with confirmatory factor analysis. A report on preliminary findings is available (OECD, 2021).

UC Consortium on Chicago School Research (UCCCSR)

UCCCSR conducted a comprehensive literature review to identify the noncognitive factors critical to children’s learning and development (Farrington et al., 2012). They sought to identify factors that were linked to academic achievement, at all ages and academic levels, that were simultaneously “not fixed traits,” but ones that could be shaped by the educational context. The nature–nurture distinction is fuzzy and often overblown, but a perception of fixed traits versus malleable competencies has influenced the research conducted in this area. UCCCSR’s literature review can be viewed as a framework for noncognitive factors. They proposed five categories of factors: (a) academic behaviors (going to class, doing homework, organizing materials, participating, and studying); (b) perseverance (grit, tenacity, delayed gratification, self-discipline, self-control); (c) mindsets (sense of belonging, growth mindset, self-efficacy, belief in the value of academic work); (d) learning strategies (study skills, metacognitive

strategies, self-regulated learning, goal setting); and (e) social skills (interpersonal skills, empathy, cooperation). There has been some empirical research evaluating this framework (Farruggia et al., 2016; Wanzer et al. 2019).

Other Frameworks

The range of noncognitive variables goes beyond the Big Five and the constructs identified thus far. Here, we discuss proposals that expand the list of potentially important constructs.

Social Attitudes

Saucier (2000, 2013) identified several social attitude dimensions (or belief system components: beliefs, attitudes, ideologies, worldviews) using a lexical approach. Respondents were given 389 descriptions of terms ending in “ism,” such as authoritarianism, liberalism, conservatism, and communism, sampled from the dictionary. They rated their agreement with the descriptions, and factor analyses resulted in four factors—(a) alpha (tradition-oriented religiousness, e.g., creationism); (b) beta (unmitigated self-interest, e.g., hedonism); (c) gamma (communal rationalism, e.g., utilitarianism); and (d) delta (subjective spirituality, e.g., reincarnationism). Saucier (2013) included additional items identifying a fifth factor, inequality-aversion (e.g., egalitarianism vs. elitism, jingoism). He found traditional religiousness correlated with Republican Party (within the United States) preference. Inequality aversion, subjective spirituality, and communal rationalism correlated with Democratic Party preference; and unmitigated self-interest correlated negatively with Openness/Intellect and Agreeableness.

Economic Preference Parameters

Almlund et al. (2011, Table 6), proposed a set of economic preference parameters—time, risk, and social preferences—based on the behavioral economics literature and argued that they overlap Big Five factors. Time preference (delay discounting) relates to one’s ability to delay gratification; risk preference refers to the amount of risk one is willing to assume; and social preferences have to do with leisure, altruism, trust, and positive and negative reciprocity. There are conceptual overlaps—risk-taking is a facet of Extraversion, time preference (punctuality vs. procrastination) is related to Conscientiousness, and trust is related to Agreeableness. A. Falk et al. (2015, 2016) administered measures of economic preferences along with the Big Five and found some overlaps; they also found correlations with national indicators such as educational attainment, savings rate, and risk behaviors (e.g., smoking).

Kyllonen (2016) proposed (but did not test) a taxonomy to summarize all these constructs. The major categories were the Big Five, generalized attitude dispositions, interests, personal beliefs, cultural and behavioral norms, social axioms, attitudes toward school, values, subjective well-being, economic preferences (time, risk, and social preferences), emotional intelligence, metacognition, creativity, collaboration, cognitive bias susceptibility, emotions, moods, and states of mind.

National Assessment of Educational Progress

Several other frameworks should be mentioned because of their importance in the field. One is for background (contextual) questionnaires for large-scale assessments, such as the National Assessment of Educational Progress (NAEP) (National Assessment Governing Board, 2013), PISA (OECD, 2017b), and other major assessments, such as TIMSS, PIRLS, and PIAAC. The National Assessment Governing Board (2013) identified priorities for NAEP contextual data collection, which were (a) NAEP reporting categories (socioeconomic status, gender, race/ethnicity, disability status, and English language status); (b) contextual factors with relationships to achievement; and (c) subject-specific information, with priority based on validity, reliability, universality (can be collected from all students), currency, respondent burden, logistic feasibility, cost-effectiveness, timeliness, nonintrusiveness, whether trends are important, and whether the information contained is valuable for understanding academic performance and how to improve it. NAEP is administered every year, and there are many different questionnaires (student, teacher, school), which are changed regularly. The student questionnaire is limited to 15 minutes. The National Center for Education Statistics (n.d.-b) provides details and the questionnaires themselves.

PISA

PISA measures reading, mathematics, and science every 3 years (since 2000; highlighting a different subject each cycle) to about half a million 15-year-olds in 70+ countries and includes a 30-minute student questionnaire. The PISA 2018 questionnaire framework (OECD, 2019b), highlighting reading, includes three construct categories: student background (e.g., out-of-school reading, socioeconomic status), schooling (teacher qualifications, instruction time, parental involvement), and noncognitive/metacognitive. This latter category includes attitudes motivation, strategies dispositional variables (achievement motives; incremental mindset; perseverance; subjective well-being; information and communications technology, motivation, and practices) school-focused variables (learning beliefs, attitudes toward school, achievement goals), and dispositions for global competence (included because global competence was assessed in the PISA 2018 survey). The latter includes communication and relationship management, knowledge of and interest in global developments, Openness and flexibility, and emotional strength and resilience. There are common themes and scales across the PISA cycles as well as changes (see OECD, 2013a, 2016, 2019b, for the past three questionnaire frameworks). There have been many analyses of the constructs together to determine the dimensionality and structure of these scales (J. Lee & Stankov, 2013; Marsh et al., 2006).

The California Office to Reform Education (CORE)

CORE is a coalition of eight California school districts serving over half a million students. It was the first U.S. jurisdiction to obtain a waiver from the U.S. Department of Education's No Child Left Behind legislation to administer a noncognitive assessment for the purposes of accountability. The questionnaire measured four constructs in

students—self-management, growth mindset, self-efficacy, and social awareness (West, 2016). There was not a systematic process for identifying these constructs, but they are some of the more popular in K–12 discussions. Two come from the CASEL five; growth mindset comes from Dweck (2017), but is related to locus of control and attribution theory, which have been a part of education discussions for decades (Graham, 1991; Rotter, 1966), as has self-efficacy (Bandura & Schunk, 1981; Schunk, 1989).

Mission Statements

Oswald et al. (2004) identified 12 critical skills for college by sorting statements coming from college mission statements. They clustered into intellectual (knowledge, continuous learning, and curiosity/artistic appreciation), interpersonal (multicultural tolerance, leadership, interpersonal, social responsibility), and intrapersonal (health, career orientation, adaptability, perseverance, ethics) behaviors. With a similar methodology, Stemler et al. (2011) identified cognitive, social, emotional, and civic as the most common themes in high school mission statements.

Competency Identification

Shultz and Zedeck (2011) conducted extensive interviews with practicing lawyers, asking questions such as, “If you were looking for a lawyer for an important matter for yourself, what qualities would you most look for?” They identified 26 effectiveness factors, divided into eight categories (cognitive, research and information gathering, communications, planning and organizing, conflict resolution, entrepreneurship, working with others, character). This approach has also been used in developing competency frameworks, which are widely used in business for promoting assessment and training needs. One example is SHL’s Great Eight (Bartram, 2005). These eight high-level competencies have been mapped to the Big Five (the letter in parentheses indicates the competency’s highest correlate (A = Agreeableness, E = Extraversion, O = Openness; N = Neuroticism; C = Conscientiousness; g = cognitive ability): learning and deciding (E); supporting and cooperating (A); interacting and presenting (E); analyzing and interpreting (g); creating and conceptualizing (O); organizing and executing (C); adapting and coping (–N); enterprising and performing (–A). Each of these is divided further into specific competencies. Leading and deciding is divided into deciding and initiating action and leading and supervising. Deciding and initiating action is further divided into six competencies, including making decisions, acting on own initiative, and taking calculated risks. Other competency frameworks are Clifton Strengths (Rath & Conchie, 2008) and Lominger’s 67 competencies (Lombardo & Eichinger, 2004), which include competencies such as dealing with ambiguity, creativity, motivating others, planning, and building effective teams. Competencies are typically identified based on factor analysis of responses to rating scale items written to represent a broad variety of attitudes, values, and behaviors relevant to workplace performance, providing an empirical foundation to this work. But there is less emphasis on finding commonalities across competency frameworks (because of market incentives) compared to the case with personality psychology.

Personality Testing Frameworks: The Big Five and Facets

Personality assessments are built around frameworks. The Big Five is an important framework. However, the most popular personality assessments typically provide many more than 5 dimensions. Analyses of items from the International Personality Item Pool (Goldberg et al., 2006), a repository of over 3,000 personality items aligned with over 400 commercial instrument scales, reveal over 25 factors (Condon, 2018). Major commercial instruments measure more than 5 dimensions: SHL's OPQ (32 dimensions); Gallup's StrengthsFinder 2.0 (Rath, 2007) (34 dimensions, called themes), and Lominger (Lombardo & Eichinger, 2004) (67 competencies). A. Costa and Kallick (2008) proposed 16 habits of mind, essential characteristics for success, including persisting, managing impulsivity, and thinking independently. C. Peterson and Seligman (2004) proposed 26 character strengths rolled up into 6 classes of virtues. A consistent finding is that the 5-factor model (or the 6-factor HEXACO model, Lee & Ashton, 2004) is the highest order of a hierarchical model (John & Srivastava, 1999; Drasgow et al., 2012). Although facets (lower order components) tend to be less stable across samples and languages than the Big Five, the 5- (or 6-) factor model is the best model we have of the major independent dimensions of human descriptions of personality, with additional structure of maybe 20 or 30 reliable factors beneath the Big Five.

Comparison of Frameworks

The frameworks reviewed only scratch the surface—Berg et al. (2017) identified 136 frameworks in the K–12 area alone! And there are higher education frameworks, such as ACT's holistic framework (Camara et al., 2015). A question is, How can such frameworks be evaluated or compared? A report from the CASEL assessment work group (Blyth et al., 2019) proposed 10 criteria. Five concerned conceptual clarity: specificity (defines competencies), balance (includes interpersonal, intrapersonal, and cognitive competencies), developmental (specifies how development occurs), culturally sensitive (accommodates cultural differences), and empirically grounded (reflects findings of associations between competencies and school, work, and life success). Five concerned implemental support: intended for practice, resources for practitioners, resources for use by children and youth, resources for measurement and data use, and empirically tested. The criteria were developed with practitioners in mind—school staff seeking a framework for implementation in response to district, state, or federal policies or mandates. Related CASEL efforts include a white paper on what frameworks are and why they are useful (Blyth et al., 2018) and tools for selecting and cross-walking frameworks (Jones et al., 2019).

The Big Five personality model has been influential in framework development. OECD's Study on Social and Emotional Skills program has adopted it, and workforce testing has embraced it. Long-standing workforce tests are issuing five-factor model conversions and score reports, and new assessments are specifically tailored around the Big Five. Why is that? The five-factor model likely has more research—reliability, prediction studies, and, increasingly, personality change studies—than any other model of individual differences in social and emotional dimensions. The Big Five has become a

standard and therefore a means to align findings and advance the field quickly. CASEL language is more popular, certainly within K–12 education, and it is highly represented in policy statements and in state social-emotional learning standards (SEL itself is a CASEL coinage). The Big Five’s power is that it emerged from an empirical finding regarding the words we use to describe ourselves and others (Goldberg, 1993). For these reasons, the Big Five model and CASEL’s five competencies are the dominant frameworks today. Few, if any, frameworks can satisfy all the requirements laid out by Blyth et al. (2019). It is likely that for the foreseeable future no one framework will be adopted universally for diverse uses. As frameworks are implemented, for various purposes, and results are analyzed, we may see more consolidation based on what kinds of frameworks prove most useful.

METHODS

Separating methods and constructs is difficult. The construct of personality is almost completely confounded with the rating scale method for measuring it. In the research literature and in policy discussions, mention of personality is invariably linked with rating scale measurement. We suspect that resistance to implementing personality assessment in schools has to do with the limitations of rating scale methodology. Despite widespread agreement on construct importance, there may be doubt about the adequacy of ratings scales for measuring them. Imagine a world in which, instead of mathematics tests, we administered self-ratings of mathematics knowledge and ability. Significant decisions would not likely be based on such data.

Disentangling method and construct is useful conceptually and for the field to advance, which has been acknowledged for over half a century (Campbell & Fiske, 1959). There may be methods other than rating scales for measuring personality. *Persistence* can be measured variously:

- Self-rating: “How often do you work on a task until you are finished: Rarely/sometimes/often/always or almost always?”
- Teacher rating: “X completes their assignments on time: True/Not true?”
- School administrative record: Missed assignments____; tardiness____
- Performance test: Length of time spent adding columns of numbers before requesting a break
- Situational judgment test: “You have a test the next day and don’t feel fully prepared. You are very tired and you are not thinking clearly. What do you do?”
- Behavioral interview: “Tell me about a time when you had to persist on a task despite many barriers in your way.”

Here, we review the main methods used to measure intra- and interpersonal skills. We review the main issues in the design, development, and scaling of assessments and the key evidence and challenges with respect to reliability, validity, and fairness (for additional information on these three concepts, see Lee & Harris; Lane & Marion; Zwick; and Rodriguez & Thurlow, all in this volume).

Self-Report Ratings

Self-report ratings are rating scale measures, known as Likert-scale measures (Likert, 1932). They refer to a statement or a question a respondent is asked to express agreement or other judgment about (e.g., behavior frequency). An example is, “How much do you agree with the statement, I am a hard worker? Strongly agree/agree/neither agree nor disagree/disagree/strongly disagree,” or “How often do you come to class prepared? Never or almost never/seldom/sometimes/often/always or almost always.” The research we have reviewed to this point is based on this method. The contextual questionnaires in NAEP, PISA, PIAAC, World Health Organization’s Health Behavior in School-Aged Children survey (Inchley et al., 2018), and most assessments in RAND’s Education Assessment Finder (Hamilton et al., 2018) and CASEL’s (2020b) SEL Assessment Guide are based on this method, as are the assessments used in the CORE Districts and the items in ETS’s SuccessNavigator. The Big Five personality model is based on self-report ratings.

There are general rules for item writing (Alreck & Settle, 1994; Fowler, 2006; National Assessment Governing Board, 2002). Items should be precise (e.g., regarding who the question refers to, the time frame), concise (e.g., ask one question, one topic; avoid “double-barreled” questions; do not use more words than necessary), neutral (e.g., avoid leading the respondent to a preferred answer), and simple (avoid esoteric vocabulary or complex sentence constructions). An easier rule of thumb is the headline rule. Imagine a *New York Times* headline that summarizes findings from the survey based on one item. “40% of New York 4th graders agree that what they learn in school is important for their future” is fairly interpretable. “30% of New York 4th graders *often believe* that what they learn is irrelevant” is probably not—what does “often believe” mean?

The NAEP contextual questionnaire group (Bertling, 2015) developed a format taxonomy for rating scale questions, applicable to questionnaire assessments broadly. Their 12-scale taxonomy specified *frequency* (“how often”) versus *amount/extent* (“how much”); frequency scales were abstract or quantified, with abstract being *absolute* (without reference point) or *relative* (with a reference point). Two types of quantified scales are absolute (i.e., counts) and *average quantified frequencies* (average counts within a time period). There are eight amount/extent scale types: time, emphasis, possibility to change, similarity to self (“like me”), agreement, confidence, likelihood, and importance. Each is associated with specific recommended response categories. An absolute abstract frequency scale uses the categories “never or hardly ever,” “once in a while,” “sometimes,” “often,” and “always or almost always.” An average quantified frequency scale uses the categories “never,” “about once or twice a year,” “about once or twice a month,” “about once or twice a week,” “every day or almost every day,” and “several times a day.” An agreement scale uses the labels described in the previous paragraph; a similarity-to-self scale uses “not at all like me,” “a little bit like me,” “somewhat like me,” “quite a bit like me,” and “exactly like me.” This was implemented in PISA 2022 (OECD, 2023).

Research issues include how many response categories to use (e.g., 2, “true of me” versus “not true of me” to 11 or more, such as a continuous graphic scale) (Revilla et al., 2014,

suggested 5 for agreement scales) and whether to include a neutral middle point (e.g., “neither agree nor disagree”) (Moors, 2008, suggested probably not). There have been studies on the interpretation of frequency quantifiers, such as seldom and often (Bocklisch et al., 2012; Newstead & Collis, 1987), which may be useful in making response category labels by spreading out the categories (selecting category labels corresponding to a spread of specific percentages that might be informative). But determining the most effective approach for a particular context is difficult. An alternative is analyzing response data after data collection to determine category frequencies and to explore the relationship between item responses and the underlying trait level using nonparametric methods, such as nonparametric option characteristic curves (Ramsay, 1991), or item response theory (IRT) methods, such as the nominal response model (Bock, 1972). R. H. Meyer et al. (2018) provided examples with the California CORE data.

Reliability

Two general approaches to analyzing and scoring rating scale data are classical test theory (CTT) and item-response theory (IRT). The CTT approach models test (or scale) scores (with exceptions; Legree et al., 2021). It assumes an observed test score is the sum of a true score and a random error. Ratings are assumed to be on an interval (continuous) response scale. A rule of thumb is that more than six ordered response category responses can be treated as continuous (Millsap, 2011, p. 121), but with fewer, they should be treated as categorical (Lubke & Muthen, 2004). Nevertheless, treating rating scale data as continuous is still the most common approach. The method involves transforming category responses (e.g., “strongly agree,” “agree,” “disagree,” “strongly disagree”) to numbers, usually 1 to k (sometimes, 0 to $k - 1$), where k is the number of categories, and then computing descriptive statistics (mean, standard deviation, skewness), and reliability (usually Cronbach’s alpha) from the continuous variable. West et al. (2020) used a CTT approach when analyzing the CORE Districts data. A self-management scale (e.g., “I came to class prepared”) used a five-category response scale (“almost never,” “once in a while,” “sometimes,” “often,” and “almost all the time”). The scale metric was transformed from the five scale categories, yielding a mean of 4.05 (presumably, the nine self-management items mean), a standard deviation of .69 and an alpha of .85 was reported (their Table 2). All four construct scales (self-management, growth mindset, self-efficacy, and social awareness) had five response scales, and although they had different labels (“not at all true” to “completely true” for growth mindset; “not at all confident” to “completely confident” for self-efficacy; “not carefully at all” to “extremely carefully” for social awareness), this allowed some degree of comparability across scales.

CLASSICAL TEST THEORY RELIABILITY In CTT, reliability is the proportion of true score to observed score variance, $r_{xx'} = \sigma_t^2 / \sigma_x^2$. Reliability is a function of both the average interitem correlation (or covariance) and the number of items in the scale, as can be seen from a formula for alpha, $r_{xx'} = \frac{n\bar{c}}{\bar{v} + (n-1)\bar{c}}$ where n is the number of items

in the scale, $\bar{c}$ is the mean interitem covariance, and $\bar{v}$ is the average item variance.³

One implication of this is that scales can be made more reliable simply by adding more items. This is shown in the Spearman–Brown prophecy formula, $r_{xx'}^* = \frac{kr_{xx'}}{1 + (k-1)r_{xx'}}$, where $r_{xx'}^*$ is the expected new reliability resulting from changing the test length by a factor of k ($k = 2$ to double the number of items or $k = .5$ to halve the number).

Reliability is important when correcting correlations with external variables and when determining effects of changes in reliability on changes in correlations with external variables. Disattenuated correlations between a scale, x , and an external variable, y , that is, correlations corrected for unreliability in both x and y , are computed as $r_{xy}^* = r_{xy} / \sqrt{r_{xx'} r_{yy'}}$. This can also be used for estimating changes in correlation with external variables, from r_{xy} to r_{xy2} based on changes in reliability, from $r_{xx'}$ to $r_{xx'2}$ is $r_{xy2} = r_{xy} \sqrt{r_{xx'2} / r_{xx'}}$; that is, the new predictive correlation is increased by a factor of the ratio of the square roots of the new to old reliabilities (Nunnally & Bernstein, 1994, eq. 7-1, p. 257).

IRT SCORING AND RELIABILITY IRT can be used to analyze and score data from rating scales. IRT treats the ratings as categorical variables and models the probability of choosing each rating category as a function of item parameters and person latent trait. There are many IRT models—van der Linden (2016) comprises 33 chapters, each dedicated to an IRT model (or class of models).⁴

There are ways to taxonomize IRT models to focus on the ones most important for the analysis of rating scales. First, distinguish continuous from categorical latent variables. Categorical latent variable models include mixture models and latent class models. These are relatively rare in analyses of rating scale data, although there are some applications for detecting response styles and faking (Eid & Zickar, 2007).

The most common IRT models assume a continuous latent variable underlying the ordered categorical responses. What is being modeled is the probability of selecting one of the k categories, such as “agree” or “strongly agree.” A set of “divide-by-total” polytomous IRT models (Thissen & Steinberg, 1986) that share common parameterization but with decreasing degree of constraints includes rating scale models (RSM; Andrich, 2016), partial credit models (PCM; Masters, 2016), generalized partial credit models (GPCM; Muraki & Muraki, 2016), and nominal response models (NRM; Bock, 1972). Item parameters for these models include item/category slope and category intercept parameters, which, respectively, represent the discrimination of the item/category as a function of the latent trait and the relative endorsement of a category.

The RSM (Andrich, 2016) assumes all item slopes are the same, and the differences of category intercepts between adjacent categories (e.g., disagree and agree, agree and strongly agree) are the same across items. The PCM assumes equal item slopes, but relaxes the constraints on category intercepts. The GPCM further relaxes the assumption of equal item slopes for all items in the scale, an assumption that characterizes

both the RSM and the PCM. However, discrimination is common across all categories. The NRM even further allows different categories to have different slopes. These are IRT models ordered by flexibility. Fitting these models to the same data can be used to test out different model assumptions (Preston et al., 2011). A cost for flexibility is the requirement for larger sample sizes. De Ayala (2009) recommended minimum sample sizes of 250, 250, 500, and 600 for the RSM, PCM, GPCM, and NRM, respectively. Once an IRT model has been determined, an individual can be scored as the estimate of the latent trait in the given IRT model.

In IRT, test information and measurement error vary with the latent trait. To get overall reliability averaging across the latent trait, Green et al. (1984) suggested marginal reliability. Following the CTT reliability definition of variance of true scores over variance of observed scores, the IRT marginal reliability is calculated as $(\sigma_{\theta}^2 - \bar{\sigma}_e^2) / \sigma_{\theta}^2$, where σ_{θ}^2 is the variance of latent trait and $\bar{\sigma}_e^2$ is the average error variance across the latent trait distribution.

CTT VERSUS IRT FOR RATING SCALE METHODS Most personality research and non-cognitive assessment research more generally employs CTT methods. CTT methods are convenient; well known; easy to apply with a wide variety of software, including spreadsheet software; and familiar to researchers and stakeholders; they can also be appropriately applied in small samples. IRT models are more complex, less well known, require specialized software, and generally require larger samples. But there are significant advantages to IRT methods. IRT overcomes the fundamental item–person confounding problem in CTT, separating item and person effects (van der Linden, 2016). This separation leads to many advantages: IRT can be used to provide better item diagnostics, scoring of test takers based on response patterns, form assembly, equating and linking of forms, assessment of change, and examination of measurement invariance across test-taking subgroups. These advantages are routinely discussed in IRT textbooks (Reise & Revicki, 2015), and the field of educational measurement has embraced IRT approaches. Because educational measurement increasingly includes personality and noncognitive assessment, a similar adoption of IRT methods in this realm can be expected.

Validity

EVIDENCE BASED ON CONTENT Rating scale items are simply statements about which a respondent expresses agreement (or frequency, amount, or extent). So, if respondents indicate that they “strongly agree” that “I am a hard worker,” this can be taken at face value as an indication of Conscientiousness, by the definition of Conscientiousness. Messick (1995) referred to this as content relevance. This *prima facie* case is bolstered by additional evidence that ratings on other indicators of Conscientiousness (e.g., “I am thorough in my work,” “I complete tasks in an efficient manner”) tend to intercorrelate. The literature on item intercorrelations based on the five-factor model attests to the nature of those relationships (John & Srivastava, 1999).

This interpretation can be challenged based on *calibration* and *faking* (or social desirability responding). The calibration challenge is that interpreting Respondent A's "strongly agree" that "I am a hard worker" as an indication of a higher level of Conscientiousness than Respondent B's "agree" that "I am a hard worker" (aside from the mapping of "hard worker" to "Conscientiousness," discussed in the previous paragraph) assumes that the two respondents have similar understandings of both the stem and the categories (Böckenholt, 2004). An example of miscalibration can be found in the meta-analytic correlation between self-estimates of cognitive ability and cognitive ability measured by tests: they are only correlated around $r = .33$ (Freund & Kasten, 2012).

The faking challenge has to do with respondents presenting themselves in a too-favorable light (self-presentation); social-desirability bias (Paulhus, 2002). Niessen et al. (2017) demonstrated self-presentation with university applicants who completed a set of ratings and then repeated the ratings after being selected into the program (reduced self-presentation). Scores were higher and outcome predictions were lower with self-presentation. West et al. (2018) minimized self-presentation by assuring anonymity, attaching no stakes to the assessment, and having the proctor (teacher) stay in the back of the room rather than roam the aisles looking over students' shoulders. Still, self-presentation presents a challenge, particularly in a high-stakes context.

EVIDENCE BASED ON CORRELATIONS WITH OTHER MEASURES The other key source of validity evidence for rating scale measures is in their correlations with external measures. Meta-analyses have shown moderate correlations between personality self-ratings and grades in higher education (McAbee & Oswald, 2013; Nofle & Robins, 2007; Poropat, 2009; Richardson et al., 2012; Robbins et al., 2004, 2009; Vedel, 2014). Conscientiousness is the highest correlate, typically in the mid .20s range. These studies, as well as the ones cited next, are typically conducted in low-stakes research settings, giving respondents little incentive to fake self-ratings.

West et al. (2018) found moderate correlations within grade between the four CORE Districts scales (self-management, growth mindset, self-efficacy, and social awareness) and grade point average (GPA; $r = .18$ to $.44$) and math ($r = .14$ to $.37$) and English language arts test scores ($r = .15$ to $.34$) (N s ranged from 7,893 to 127,134). This aligns with expectations based on our review and supports a test score interpretation that the four constructs are associated with educational success.

There are two key challenges to this interpretation, independent of the challenges associated with the calibration and faking issues discussed previously. The four scores are correlated quite highly, relative to their reliability. At the school level, correlations range from $r = .36$ to $r = .97$, with a median $r = .91$ (Hough et al., 2017). This suggests that there might not be four unique pieces of information in the four scale scores, which would challenge an interpretation based on the scale content (e.g., self-efficacy, growth mindset), perhaps toward an alternate interpretation based on the common elements across the scales (e.g., social desirability). The fact that West (2016) showed that the highest correlate of English language arts grades was the composite of all four measures,

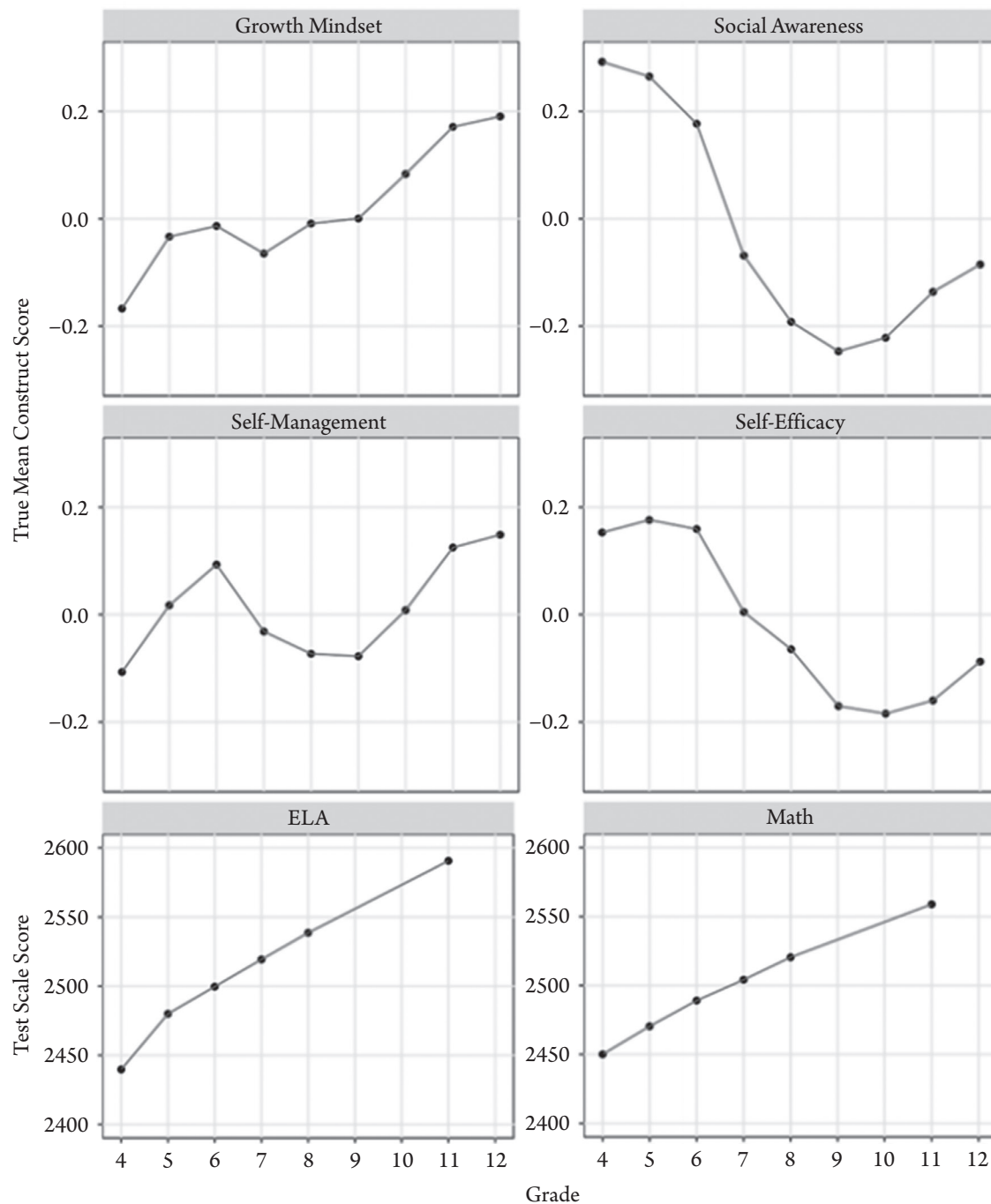
rather than any individual measure, supports this idea. A subscore analysis, as suggested by Haberman (2008), would be informative to establish scale independence.

Second, interpreting results across grades is challenging. Figure 19.1 shows the cross sectional comparisons of CORE District students' SEL scores from Grades 4 to 12 and, for comparison purposes, the math and English language arts achievement test scores. Achievement test scores, which are vertically scaled and therefore comparable, are monotone, almost linear increasing. The SEL scores are from identical questions (i.e., the same questions were asked in the 3rd through 12th grades), scaled to a common metric. They sometimes go up, sometimes down, without a clear pattern. At the least, this challenges the interpretation of test scores as indicating an association of the four constructs with achievement. Fourth graders' social awareness is considerably higher than that of high schoolers, yet their achievement is considerably lower. Within a grade, the correlation is higher. The cross-grade pattern is difficult to interpret. Generally, age, period, and cohort effects all might influence findings like these (Ion et al., 2022).

Fairness

Fairness in assessments as well as communications about those assessments, in various standards documents (American Educational Research Association [AERA] et al., 2014; ETS, 2014, 2022), refers to the requirement that construct-irrelevant personal characteristics of test takers have no appreciable effect on test results or their interpretation. Fairness reviews conducted prior to pilot testing consider representation of the groups to be studied, accommodations, presentation and response formats, the potential effect of timing (e.g., time limits), language (e.g., English language learning status), experience with the technology, and prohibition of inappropriate language, among other factors. Noncognitive and cognitive assessments are not different in these regards. Generally, the goal is to avoid materials that are not fair because they include the wrong content and skills or they fail to include a good sample of the right content and skills (ETS, 2022).

MEASUREMENT INVARIANCE Measurement invariance analyses is conducted to ensure comparability of interpretation across groups (e.g., race/ethnicity, gender, age, language) or over time. With continuous data there are well-documented approaches to establishing measurement invariance through a sequence of tests for configural, metric, and scalar invariance. Tests for configural invariance establish that the factor structure is the same across groups; metric invariance establishes that item loadings are the same; scalar invariance establishes that the item intercepts and loadings are the same. Establishing metric invariance allows one to interpret the correlations between scale scores and external variables as comparable for the different groups; establishing scalar invariance allows one to interpret the mean scale (factor) differences between groups. Failure to establish measurement invariance at any point makes comparisons between groups problematic. It is not appropriate to claim a mean "Conscientiousness" difference between males and females if scalar invariance is not established. If a Conscientiousness item was "I enjoy assembling and disassembling car engines" and the mean difference on

**FIGURE 19.1**

Cross-Sectional Comparisons of CORE District Students Scale Scores Across Social-Emotional Learning (SEL) Scales (Top Four Panels) and Achievement Test Scores (Bottom Two Panels)

Note. Achievement test scores are vertically scaled and monotone increasing. SEL scale scores are from identical questions across years scaled to a common metric. They are not monotone increasing. From *Trends in Student Social-Emotional Learning: Evidence from the CORE Districts* (Working Paper), by M. R. West, L. Pier, H. Fricke, H. Hough, S. Loeb, R. H. Meyer, and A. B. Rice, 2018, PACE: Policy Analysis for California Education, CORE-PACE Research Partnership. Reprinted with permission.

a Conscientiousness scale that included the item showed a higher mean for males, one might be reluctant to conclude that males were more conscientious than females (Ployhart & Oswald, 2004).

With categorical data, as typically found on rating scales, the configural–metric–scalar series of tests is not appropriate, unless those data are treated as continuous. But if they are treated as categorical, an alternative sequence of first testing for configural invariance, then for threshold invariance, then for loadings and intercept invariance is recommended (Svetina et al., 2020; Wu & Estabrook, 2016). There are several reasons for conducting measurement invariance analyses. One is to determine the appropriateness of comparing groups. Another is to explore reasons for misfit, which can provide additional insight into group differences (Putnick & Bornstein, 2016). Typically, when conducting measurement invariance tests, steps are taken to identify misfitting items and delete them or relax certain assumptions about them (e.g., thresholds). This process can provide insights on possible reasons for misfit and allow for comparison of scale scores between groups. The practical importance of lack of invariance should also be considered (Nye & Drasgow, 2011; Schmitt & Ali, 2015).

Response Style

For rating scale items, response style is potentially a source of construct-irrelevant variance. Response style is a major factor in international assessments involving different languages, but it also can also differentiate subgroups within a language group (e.g., gender, race/ethnicity, immigrant status). There are well-documented differences between cultural groups in using the response scale, independent of the construct being measured. These include extreme response style (tendency to choose scale extremes, such as “strongly disagree” or “strongly agree”), modest or midpoint response style (tendency to choose the scale midpoint, such as “neutral” or “neither agree nor disagree”), and yeah-saying or acquiescence response style (ARS; tendency to agree). Response style effects can be adjusted for statistically. Khorramdel and von Davier (2014) identified extreme response style and modest or midpoint response style using a multinomial processing tree approach, showing the generalizability of response styles across response scales. He and van de Vijver (2015), using Teaching and Learning International Survey data, interpreted a general response style factor with positive loadings from extreme response style and socially desirable responding and negative loadings from modest or midpoint response style as response amplification versus moderation. Primi et al. (2019) identified positive and negative key item pairs (close to opposite) within a scale and presented both classical and IRT corrections to bias (typically positive bias, such as responding “strongly agree” to one of the pair and “neutral” or “disagree” to its opposite). C. F. Falk and Cai (2016) provided a general approach to modeling response styles. A special issue of the *British Journal of Mathematical and Statistical Psychology* (Khorramdel et al., 2019) assembled eight different approaches for correcting rating scale data for response style, including multidimensional IRT, IRTree, and clustering approaches. Some of these methods should routinely be applied to rating scale questionnaires.

Although rating scale approaches, particularly self-ratings, are by far the most common in the assessment of interpersonal and intrapersonal skills, there are significant challenges in their interpretation as construct measures. Also, while self-rating scales can be used in low-stakes assessments, they are problematic for high-stakes use (Niessen & Meijer, 2017).

Others' Report Ratings

Others' (informant) ratings refers to evaluations by others—peers, supervisors, professors, or others who are in position to evaluate the target. Unstandardized evaluations, such as letters of recommendation, are common, particularly in higher education and the workforce (personnel selection, promotions), but little research exists on them. Despite that, letters of recommendation and associated ratings are considered quite important in education admissions (Kuncel et al, 2014) and in the workforce (Judge & Higgins, 1998). Perhaps the reason for their popularity is that they protect against the faking and potential deception that can characterize self-reports. Personal statements are commonly required in admissions and scholarship competitions, despite little evidence for their value (Murphy et al., 2009). The reader of a recommendation letter can be at least somewhat assured that a relatively disinterested party is endorsing a candidate.

Research on letters of recommendation is based on post hoc numerical summaries of letter contents (Baxter et al., 1981; Kuncel et al., 2014). Such summaries correlate with outcome measures in higher education, GPA, and degree attainment and provide small, but incremental prediction beyond other measures in predicting attainment (Kuncel et al., 2014). But a quantitative summary of letter contents is not routinely available and is not how letters are typically used, which is more anecdotal. The future may bring natural language processing analyses of letters, as with essays and constructed response tasks (Madnani & Cahill, 2018), but little of that has appeared so far.

Many formats exist for others' ratings besides open-ended letters of recommendation. They range from a standardized letter of recommendation used for graduate admissions, such as ETS's Personal Potential Index (Kyllonen, 2006), to a clinical psychologist's rating of candidates' psychological fitness for military duty based on a 30-minute interview (Lindqvist & Vestman, 2011), a teacher's checklist of classroom misbehavior and a parent questionnaire used in the National Educational Longitudinal Study of 1988; National Center for Education Statistics, n.d.-a), and teacher and parent rating forms for students found in RAND's Education Assessment Finder (Hamilton et al., 2018) and CASEL's (2020b) SEL Assessment Guide, as well as for large-scale assessments such as the PISA Parent Questionnaire (OECD, 2017b), and the CORE Districts Teacher Questionnaire (West et al., 2018).

It is useful to consider the constructs and measures used. ETS's Personal Potential Index comprised six scales with a text box for each scale, which allowed evaluators to justify or elaborate on the ratings. The six scales were knowledge and creativity, communication skills, teamwork, resilience, planning and organization, and ethics and integrity.

Each scale comprised four items. For planning and organization, the items were sets realistic goals, organizes work and time effectively, meets deadlines, and makes plans and sticks to them. The rating scale included six categories: “below average,” “average,” “above average,” “outstanding (top 5%),” “truly exceptional (top 1%),” and “do not have sufficient information to evaluate.”

The Swedish military interview (Lindqvist & Vestman, 2011) was based on a clinical psychologist’s analysis of several attributes, combined to form a single composite score. The positive attributes were willingness to assume responsibility, independence, outgoing character, persistence, emotional stability, initiative, and social skills. Also considered was ability to adjust to the specific requirements of life in the armed forces (loss of personal freedom). Specific traits screened out were difficulty accepting authority, violent or aggressive behavior, psychopathology, and antisocial tendencies.

The CORE Districts teacher ratings were only for the self-management and social awareness scales, presumably because self-efficacy and growth mindset are largely unobservable. There were nine self-management questions and eight social awareness questions. The teacher questions were in some cases nearly identical to the self-assessment versions (student self-assessment: “I remembered and followed directions” versus teacher rating: “remembered and followed directions”), but in some cases there were small discrepancies (student self-assessment: “I was polite to adults and peers” versus teacher rating: “got along with others”). The response categories for the two scales were the same and were the same for teachers and students (“almost never,” “once in a while,” “sometimes,” “often,” “almost all the time”).

Others’ ratings have several conceptual advantages over self-ratings. An external other can place an observed target (a student) in a distribution of other observed targets (other students), whereas a self-rating is subject to biases discussed in the previous section (e.g., social desirability). Also, two or more independent raters can rate a target, whereas only one self exists (there can be different imagined selves, as in faking experiments).

Analysis, Scoring, Scaling, and Reliability

Almost all analysis, scoring, and scaling of ratings that appear in the literature is based on simple sum scoring across raters (R_1 , R_2) and classical test theory for evaluating ratings. Often, a statistic to measure rater agreement is computed, such as percent agreement, or Cohen’s (1960) kappa, which adjusts for agreement due to chance. This can be generalized to the situation with multiple categories (as with Likert scales), where closeness is credited using the weighted kappa statistic (Cohen, 1968), and with multiple raters (Conger, 1980). Although these methods remain popular, Gwet’s (2014) AC1 and AC2 statistics may be useful alternatives in some cases (Keener, 2020).

Generalizability theory (Brennan, 2001; Cronbach et al., 1972; see also Lee & Harris, this volume) can be used for analysis of ratings data. Generalizability theory is an extension of classical test theory in which measurement error is not unitary but is partitioned into sources such as multiple items, occasions, raters, or other factors. A generalizability theory analysis was conducted on the Personal Potential Index by McCaffrey et al.

(2018). They found high correlations among the six scales ($r = .67$ to $r = .90$), consistent with a halo effect (Thorndike, 1920). They found a large rater effect, larger than the effect for persons being rated. This is consistent with research (Baxter et al., 1981) finding more similarity between two ratees rated by the same rater than between two rated by different raters. Consequently, McCaffrey et al. (2018) estimated that 10 raters or possibly more might be needed to obtain a score that had the same level of reliability as commonly used admissions assessments. The authors also point out that there can be response style effects operating on evaluators, as there are on self-raters, including severity and leniency tendencies. These factors (halo, lack of rater agreement, rater response biases) likely operate in all rating situations but are only apparent when ratings are quantified.

IRT approaches can evaluate raters, most notably with the Rasch-based Facets software (Linacre & Wright, 2002), which has similar purposes as a generalizability theory analysis. Facets is designed to analyze data from multiple items and raters and estimates rater leniency-severity effects. To our knowledge, Facets has primarily been applied to ratings of cognitive tasks (e.g., writing), rather than to Likert-style ratings of noncognitive competencies.

Validity

EVIDENCE BASED ON CONTENT The content analysis case for ratings by others is essentially the same as that for self-ratings—ratings are straightforward statements expressing the rater's belief about the target based on observations. Others' ratings better predict outcomes than self-ratings with traits high in evaluativeness, such as intellect (Vazire, 2010). The key challenges to self-ratings, calibration and faking, are generally less applicable to observer ratings. An observer can compare the target to others without self-serving bias, increasing calibration, and the observer typically has minimal incentives to fake (there may be special circumstances where this is not true, such as the professor who takes pride in their students' accomplishments). Leniency severity and halo bias exist. Also, calibration is not guaranteed (McCaffrey et al., 2018). Some not-easily-observable traits are known only by the self (Vazire, 2010).

EVIDENCE BASED ON CORRELATIONS WITH EXTERNAL VARIABLES Several meta-analyses have evaluated others' ratings for predicting school (Connelly & Ones, 2010; Poropat, 2014) and workforce outcomes (Connelly & Ones, 2010; Oh et al., 2011). All have shown that ratings by others have higher correlations with outcomes than do self-ratings. Poropat (2014) examined Big Five factors' prediction of GPA. Targets ranged in age, but were mostly college students. The Big Five was measured by various instruments under various design conditions (one versus multiple raters; teacher vs. peer vs. spouse ratings). The main finding was that the correlations with performance given by others' ratings were higher than those given by self-reports for all of the Big Five factors (other rating $\rho = .38, .28, .18, .10, .05$ versus self-rating $\rho = .22, .09, .00, .06, -.02$ for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively; Openness correlations go down substantially

when cognitive ability is controlled, but the others do not change). Poropat (2014) claimed that the magnitude of these correlations puts them among the largest effect sizes in education considering Hattie's (2009) meta-analysis of effect sizes in education. Feng et al. (2022) similarly found that teacher reports were more reliable than children's self-reports and more predictive than self-reports of school outcomes controlling for cognitive ability. These studies are generally low-stakes applications, in which ratings, by teacher, guardian, or self, are for school monitoring or program evaluation. McAbee and Connelly (2016) provided a framework for understanding discrepancies between rating sources.

Connelley and Ones (2010) also included school outcomes in their meta-analysis of mostly workplace outcomes. They found that simple correlations between self- and others' ratings of outcomes were similar, but others' ratings were substantially greater than and incremental to self-ratings in predicting outcomes when unreliability was corrected for. In addition, they conducted a meta-analysis on interrater agreement and found, like McCaffrey et al. (2018), that many raters, up to 10, would be needed to get reasonable interrater reliability; at the very least, they recommended having more than 1 rater.

Oh et al.'s (2011) meta-analysis focused on workplace outcomes. They found predictions comparable to the other meta-analyses. The true-score correlations between predictors and outcomes with other reports were $\rho = .37, .26, .21, .31, .27$, and for self-reports they were $\rho = .22, .05, .14, .10, .09$, for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively. Other report correlations were like Poropat's (2014) estimates, except for the latter two (Agreeableness and Extraversion), which were higher in Oh et al. (2011) (studies did not overlap because Poropat, 2014, used educational outcomes, whereas Oh et al., 2012, used workforce outcomes). Oh et al. (2012) showed that multiple raters improved prediction (e.g., for Conscientiousness, $\rho = .32$ and $.41$ for one and three raters, respectively); however, even correlations with outcomes based on single raters were substantially higher than those based on self-ratings. Oh et al. (2012) also conducted regression analyses to show that others' ratings added incremental R^2 prediction to self-ratings (on average, $\Delta R^2 = .048$) but that the reverse was not true (on average, $\Delta R^2 = .005$). An earlier meta-analysis (Harris & Schaubroeck, 1988) estimated correlations of performance ratings between self-supervisor ($\rho = .35$), self-peer ($\rho = .36$), and peer-supervisor ($\rho = .62$), indicating the self is the outlier.

Two other studies from the economics literature similarly reported high correlations with outcomes based on others' ratings. The Swedish enlistment study (Lindqvist & Vestman, 2011) examined earnings, employment, and chronic unemployment 20 years after an overall noncognitive rating by one clinical psychologist was collected from 18-year-old males (the rating factors are discussed in a previous section). They found that the psychologist's noncognitive rating was a stronger predictor of outcomes than cognitive ability, particularly for the outcome of chronic unemployment, where it was a stronger predictor than cognitive scores by a factor of 5.

The National Educational Longitudinal Study of 1988 teacher ratings study (Segal, 2013) examined earnings 20 years after two eighth-grade teachers' ratings of student misbehavior (dummy variable: 1 if either teacher rated "yes" to any of the items: *rarely completes homework, is frequently absent, is frequently tardy, is consistently inattentive in class, and is frequently disruptive*; 0 otherwise). She found that after controlling for test scores, eighth-grade misbehavior predicted earnings, and after controlling for educational attainment, it predicted earnings at all educational levels, whereas achievement predicted earnings only for males with postsecondary degrees.

A question is, How much do others' ratings correlate with self-ratings? West et al. (2018) presented correlations between teacher and student self-ratings of $r = .41, .38$, and $.47$ for elementary, middle, and high school students, respectively, for self-management and $r = .22, .21$, and $.26$ for social awareness (teacher ratings were averaged when there was more than one teacher). However, correlations between teacher ratings on the two scales were considerably higher, $r = .85, .82, .85$, for elementary, middle, and high school teachers, indicating a considerable halo effect! Connelly and Ones (2011) also found that self-other agreement was at least moderately high. The mean observed self-other correlations were $\bar{r} = .37, .34, .34, .29$, and $.41$, for Conscientiousness, Openness, Emotional Stability, Agreeableness, and Extraversion, respectively, boosted substantially when correcting for unreliability. West et al.'s (2018) self-management and social awareness are similar to Connelly and Ones's (2011) Conscientiousness and Agreeableness, respectively, and then the correspondence in findings is much clearer. Also, Connelly and Ones (2011) found that the most accurate others' ratings came from friends and family (they did not examine teachers as a separate category); but there were some traits, such as Extraversion, for which strangers' ratings were as accurate as ratings from those who had known the target longer.

Fairness

The use of differential item functioning and measurement invariance approaches for analyzing ratings data is extremely rare, partly reflecting the rarity of ratings data for noncognitive characteristics generally. Also, response style effects and social desirability could undoubtedly be found in ratings data, but they have not been systematically investigated. Two other rating biases have been examined. One is the halo effect (Thorndike, 1920), the lack of differentiation in ratings across constructs. A halo effect was observed in the results of West et al. (2018), as the extremely high intercorrelations among ratings on the four CORE Districts variables (all $r > .80$). Similarly, McCaffrey et al. (2018) showed very high correlations among the six Personal Potential Index ratings (median $r = .81$). The other bias is rater leniency versus severity, which is like acquiescence response style in self-ratings. We are not aware of a body of literature that has examined this phenomenon in noncognitive ratings, although it is well studied and documented in ratings of cognitive test performance (Linacre & Wright, 2002).

In summary, others' ratings likely overcome some problems with self-ratings, most notably the reduction in social desirability, faking, response style, and calibration

biases. This may be what is responsible for findings of increased prediction of outcomes compared to self-ratings, which was found in several meta-analyses of both school and workforce outcomes. However, others' ratings introduce some additional problems, particularly halo, which results in a lack of differentiation between constructs, and the low level of agreement between raters. A common recommendation is to include others' ratings where possible, particularly more than one rater, with an awareness of their limitations.

Anchoring Methods

Anchoring methods refer to a set of techniques for increasing rating scale comparability across targets using scale anchors. These methods are intended not to mitigate or adjust for faking, but to increase interpretability of responses, particularly in low-stakes assessments. A scale anchor is a concrete behavioral descriptor or vignette, which may be more resistant to response biases than response category labels like "agree" or "often." Examples of anchoring methods are anchoring vignettes (AVs; King & Wand, 2007), behaviorally anchored rating scales (BARS; Kell et al., 2017; Klieger et al., 2018), behavioral summary scales (BSS; Borman, 1979), Rasch/Guttman scenario scales (Ludlow et al., 2020), and perhaps single-item measures of personality, bipolar single-item scales with richer descriptors (Woods & Hampson, 2005), which could have been included here, but we do not discuss them in detail. In principle, all of these can be used as self or other ratings, although AVs and single-item measures of personality are typically used as self-ratings and BARS, BSS, and Rasch/Guttman scenario scales are typically others' ratings.

Anchoring Vignettes

AVs supplement self-rating scale measures with additional vignettes that describe one or more hypothetical persons—the respondent's task is to rate vignettes the same way they rate themselves (on the same rating scale, with the same stem). With more than one vignette, the vignettes are designed to be at different locations on the scale (e.g., a high and low location). Assuming vignettes are interpreted the same way by all respondents (vignette equivalency assumption) and that the rating standards for self and vignettes are the same (response consistency assumption), self-ratings can be interpreted in relation to the vignette ratings to create a new anchored self-rating scale. Consider a case with one vignette ($J = 1$) describing a hypothetical person, X:

X sometimes arrives late to class; usually, but not always completes their homework; and sometimes procrastinates on assignments. How much do you agree that X is hardworking? (strongly disagree, disagree, agree, strongly agree).

The respondent is also asked to rate themselves on the same item:

How much do you agree that you are hardworking? (strongly disagree, disagree, agree, strongly agree).

The two responses enable transforming the Likert self-rating to a new relative (to the vignette) rating called C_i , for each respondent i . Given an item and the self- and vignette rating on that item, the respondent might rate themselves higher, lower, or at the same level as they rated the vignette; thus, the new scale C_i is a three-category scale. With two vignettes ($J = 2$), there are more possibilities (rate self higher, the same, or lower than the high-location vignette and higher, the same, or lower than the low-location vignette), and with three vignettes, still more (the number of categories in C is $2J + 1$). Response ties (rating the vignettes the same) and misorderings (rating the lower of two vignettes higher than the higher vignette) reduce the number of transformed categories (e.g., higher than both, lower than both, or neither), but do not entirely invalidate the nonparametric transformation rule logic.

SCORING AVs can be scored in two ways. A nonparametric approach transforms self-ratings to a new relative-to-anchor scale, as described in the previous paragraph. Some self-rating responses cannot be treated in a straightforward way because of ties or misorderings. Vignettes are written to be at a particular location (e.g., high and low), but they can be moved after the fact based on empirical findings as long as they are ordered the same for all respondents. Self-rating responses under these conditions can be treated as missing or by various assignment rules, which can be specified (Wand et al., 2016).

Parametric AV scoring uses ancillary variables such as country, gender, and race/ethnicity in addition to self- and vignette ratings to estimate scores on the latent factor, the unobserved perceived level. The parametric approach has several advantages over the nonparametric approach: Not all respondents have to rate all the vignettes, and misorderings are assumed to result from random error. The R package (R Core Team) “anchors” (Wand et al., 2011) handles both nonparametric and parametric scoring and includes many options and diagnostics.

RELIABILITY Internal consistency reliability estimation with AVs is not straightforward. Von Davier et al. (2018) showed that Cronbach’s alpha is not an appropriate measure of internal consistency for AVs, particularly when ties and misorderings occur, and that scale intercorrelations will increase when corrected by the same set of vignettes. However, test–retest correlations would still be useful (with the same or different sets of vignettes), as would a generalizability theory (variance component) analysis using vignettes as a random factor (e.g., using two sets of vignettes for the same construct; on one or multiple occasions). To our knowledge, such research has not yet been reported.

VALIDITY AND FAIRNESS The justification for AVs is that they adjust responses for response style biases and reference group effects with information from responses to the vignettes. This assertion depends on the vignette equivalency (different respondents see the vignettes the same way) and response consistency (respondents apply the same standards in rating themselves and the vignettes) assumptions. These can be vio-

lated (Kapteyn et al., 2011), and methods for evaluating these assumptions are needed. Despite this, a major finding with AVs is that they have been found to address the “attitude–achievement paradox” and increase country comparability on certain noncognitive scales (Kyllonen & Bertling, 2014). The attitude–achievement paradox is the common finding of a positive attitude–achievement correlation within country (e.g., between a personality factor and achievement), with a simultaneous negative correlation at the country level (e.g., the countries with high average personality scores will have low average achievement scores). This is a recurring finding in large-scale international assessments for many (but not all) scales, particularly ones that are more abstract, such as personality scales. Kyllonen and Bertling (2014) and He et al. (2017) showed on PISA 2012 that AV-adjusted scores on two scales (teacher support and classroom management) had higher correlations with mathematics achievement than unadjusted scores. He et al. (2017) also found that Asian cultures topped the ranking on teacher support after adjustment, a finding more in line with the literature. AV approaches have now been tried in numerous studies, with mixed results (Möttus et al., 2012; Primi et al., 2016). It is a complex method and there are many issues to consider in vignette and study design, but the findings from PISA 2012 seem promising and worth pursuing.

Rasch/Guttman Scenario Scales

Rasch/Guttman Scenario Scales is a method developed by Ludlow and colleagues (Ludlow et al., 2019, 2020), which also use vignettes, called scenarios. The respondent is presented a series of scenarios and asked to rate themselves (on a five-category scale going from strongly higher to strongly lower) relative to the character described in the scenarios; the scenarios describe people at different locations on the construct. Scenarios are developed using Guttman’s facet theory and mapping sentence approach (Guttman & Levy, 1982). This differs from AVs in that with AVs, the respondent rates self and vignette independently. Here, one compares oneself directly with the vignette (scenario). Constructs are defined hierarchically, so goal orientation might be a higher order construct, with clarity, effort, and frequency as lower order facets (analogous to the facets–factor relationship in the five-factor model; John & Srivastava, 1999). The Sentence Map contains sentences with location fillers representing different construct levels. Ludlow et al. (2020; Table 1b) provided an example:

clarity: “<name> has a {not at all clear| somewhat clear| extremely clear} vision of how to make goals a reality”

effort: “he/she places {almost no effort |some effort |a tremendous amount of effort} toward making long term aims a reality”

frequency: “he/she {almost never |sometimes |almost all the time} spends time in the day engaged in activities that bring him/her closer to his/her goals”

Because there are three mapping sentences with three locations each, there are m^n (m levels, n sentences) possible sentence combinations (e.g., LLL to HHH), but only enough need to be written (Ludlow et al., 2020, wrote 7 of the 27) to ensure spread

across construct levels. The sentences are transformed to scenarios through editing for flow and naturalness; some facets might be combined in a single sentence. Ludlow et al. (2020) (Study 2) presents six goal orientation scenarios: the highest (HHH) is “Bill knows how to make his goals a reality and constantly exerts tremendous effort toward accomplishing them”; the next highest (HHM) is “Jill is sure she knows how to achieve her goals. She is actively engaged in efforts to make her goals a reality and wishes she could be even more engaged.” The second to lowest level scenario (LML) is “Jim is unclear on how to make his plans a reality. Although he is rarely engaged in activities that move him closer to his goals, he does put some effort into working toward them.” Pilot testing can result in rewriting or swapping out scenarios (e.g., an LLH for an LHL).

SCORING AND RELIABILITY Ludlow et al. (2019, 2020) used the Rasch rating scale model for scaling and scoring. The analysis provides item and person locations on the variable maps, which can be examined to ensure good coverage on the construct continuum; revisions to items or to the selection of items can be made to ensure good coverage, as Ludlow et al. (2020) demonstrated. They also showed good model fit for their measure.

VALIDITY AND FAIRNESS The justification for using this assessment to draw inferences about an individual’s trait level follows the basic logic of other self-ratings measures—an individual is making claims about themselves. The claim is bolstered compared to simple self-ratings in that the self-description is in relation to a hypothetical other, although here, the claim is perhaps more direct than AVs; respondents are not rating themselves, then a vignette; respondents rate themselves directly compared to the vignette. This would seem to make the response consistency assumption more likely to be satisfied. A finding of good model fit supports the justification for its use. Poor model fit would indicate potential inconsistencies in the judgments being made by the respondent. Like AVs, scenario scales should reduce response style biases and reference group effects, although in Ludlow et al.’s (2019, 2020) application the rating scale (strongly higher to strongly lower) could allow response style effects.

The method is relatively recent, and little work has been done thus far on group differences or measurement invariance. However, measurement invariance approaches for Rasch measurement are available (Millsap, 2011; Wu & Estabrook, 2016).

BARS/BSS

BARS (Kell et al., 2017; Klieger et al., 2018; Oswald et al., 2004; Shultz & Zedeck, 2011) are discrete category rating scales (e.g., 0 to 5) with behavioral descriptions based on critical incidents appearing at various points along the rating scale continuum. BSS are similar, but represent composites of several incidents, resulting in more abstract behavioral descriptions. The behavioral descriptions, or anchors, help raters calibrate their ratings. The rater can locate the target relative to the anchors by essentially doing mental paired comparisons (e.g., “*x* is higher than Anchor 1,

but not as high as Anchor 2). Computer-adaptive versions of BARS, called CARS, have also been developed and evaluated with some success (Darr et al., 2017). A related alternative, the relative percentile method (Goffin et al., 2009), asks raters to rate employees against other employees, so that employees themselves serve as scale anchors.

BARS are developed through the initial collection of critical incidents, settings in which the construct is expressed through behavior. For example, Klieger et al. (2018) asked subject matter experts to recall examples of “highly ineffective, just good enough, and highly effective behavior” on job domains such as initiative/work ethic and flexibility/resilience. Subject matter experts are asked to describe the situation, the behavior, and the outcome associated with the critical incident and sometimes an interpretation (Hall et al., 1995). Incidents are sorted by dimensions or themes and then edited into behavioral statements such as “prior to transitioning to a new department, reaches out to relevant coworkers to inquire about strategies for a new position” (Klieger et al., 2018). Incidents are then rated on a six-point effectiveness scale resulting in a BARS. This general method was applied to lawyering (24 job domains; Shultz & Zedeck, 2011).

Scoring and Reliability

The fielded BARS provide ordered categorical data, typically treated in classical fashion, using sum scores and classical reliability theory. The method naturally lends itself to a generalizability theory analysis, with raters, items, settings, and occasions potentially serving as design facets over which to evaluate generalizability (Medvedev et al., 2019; Ohland et al., 2012). To our knowledge, except for Darr et al. (2017), few efforts have developed IRT models for the BARS development or application cycle.

Validity and Fairness

As an outcome measure, BARS do not have the correlational evidence that typically comes with predictor measures. Instead, for BARS, like licensure tests, validity evidence tends to be associated with the process by which the assessment is developed and based on content matches between the assessment items and the intended constructs (i.e., evidence based on test content). The BARS development process (Kell et al., 2017; Klieger et al., 2018)—the identification of the construct(s) based on job analyses, the collection of critical incidents, effectiveness evaluations, and content evaluations of items—is rigorous and designed to ensure validity evidence. Debnath et al. (2015) argued that BARS remain popular appraisal instruments because of their advantages over instruments that focus on traits (because traits are relatively immutable and subject to halo effects), results (because hard criteria are unavailable for many jobs), and behavior (such as checklists; they are too rigid and leave out evaluator judgment).

Neither differential item functioning nor measurement invariance analyses tends to be performed on BARS. However, self–other similarity bias exists for ratings. Supervisor–subordinate racial differences are associated with lower ratings for both Black and

White subordinates (Elvira & Town, 2002) and sometimes gender similarity effects are found, although not universally (Stone et al., 2016).

Forced-Choice and Ranking Methods

Forced-choice and ranking methods refer to tasks in which respondents are asked to choose from, rank, or express preferences between two or more statements. Ranking is an alternative response format to ratings. A forced-choice example would be, “Choose the statement more like you: ‘I enjoy working with others,’ or ‘I am a hard worker.’” A ranking example would be, “Rank the following 1 to 4 for how they describe you (1 = *most like you*; 4 = *least like you*): ‘I enjoy working with others,’ ‘I am a hard worker,’ ‘I am calm even in stressful situations,’ ‘I enjoy speculating about the origins of the universe.’”

Design Issues

UNIDIMENSIONAL VERSUS MULTIDIMENSIONAL FORCED CHOICE Forced-choice items may be unidimensional or multidimensional. Forced-choice statements on a unidimensional item are drawn from the same dimension but vary in their location: “I am a hard worker” versus “I do enough to get by.” The Myers–Briggs Type Indicator (Form M) uses 93 *unidimensional* forced-choice items (statement pairs), each asking individuals whether they are better described by positive or negative pole descriptors for the four bipolar factors they measure (extravert–introvert, sensing–intuition, thinking–feeling, judging–perceiving; roughly 23 pairs per factor). SHL’s Operational Personality Questionnaire (OPQ32r; SHL Group, 1999; Brown & Bartram, 2009) uses *multi-dimensional* forced-choice items (three-statement triads) to measure 32 dimensions, with items such as “Please choose one MOST true and one LEAST true statement: ‘I enjoy the companionship of others,’ ‘I try out new activities,’ ‘I look to the future.’”

BLOCK SIZE Block size refers to the number of statements ranked, typically either two, three, or four, with exceptions (Dueber et al., 2019, used 12). SHL’s two OPQ versions differ by having either three (104 triads; OPQ32r) or four (104 tetrads) statements (OPQ32i) per block (Brown & Bartram, 2013; the OPQ32r’s triads are the same as the OPQ32i tetrads with one statement removed per block). ETS’s FACETS (12 to 15 dimensions, 104–120 pairs; Naemi et al., 2014), DoD’s TAPAS (22 dimensions; Drasgow et al., 2012), and the Myers–Briggs Type Indicator Form M all use pairs.

RESPONSE FORMAT Common response formats are PICK (choose the option “most like me”), MOLE (choose the option most like me and the one least like me), and RANK (rank options 1 to n) (Hontangas et al., 2015). Some formats allow a graded preference, such as strongly prefer i , slightly prefer i , prefer neither i nor k , slightly prefer k , or strongly prefer k (Gallup’s Clifton StrengthsFinder 2.0, Asplund et al., 2014; 177 pairs to measure 34 dimensions), which can be modeled using IRT (Brown & Maydeu-Olivares, 2018). Another format is compositional (Aicheson, 1982; Brown,

2016), in which respondents express choices by dividing (say) 100 points across a set of four descriptors, such as assigning 50 to dependable, 20 to curious, 20 to modest, and 10 to calm; or selecting 10 descriptors from a pool of 30.

RESPONSE PROCESS (DOMINANCE VERSUS UNFOLDING) For a dominance response process, the higher a respondent is on a trait, the more likely they will endorse the trait (e.g., respond “strongly agree” or “almost all the time”). This is the typical assumption in personality assessments; it enables data to be fit by CTT models and the same IRT models used in cognitive testing, where the higher the ability, the more likely one is to get the item correct. An unfolding response process involves an ideal point in which a respondent is most likely to endorse an item whose location is the same as the respondent’s latent trait position and increasingly less likely to endorse an item farther away from their trait location, allowing rejection “from above” or “from below” (Chernyshenko, 2003; J. S. Roberts et al., 2000). An introvert would reject the item “I enjoy talking to a friend in a quiet cafeteria,” preferring no social interaction, and an extravert might reject the same item, preferring a more intense social experience. Not common in personality, ideal point models are common in attitude measurement, allowing disagreement with a middle-of-the-road policy from the ideological left or right. The appropriateness of unfolding for personality measurement is open to debate (Drasgow et al., 2010; Oswald & Schell, 2010). Nevertheless, using simulated and real data, Fu et al. (2022) found support for the theoretical claim but found mixed evidence with real data. They recommended fitting different models to different statements.

DATA TYPE Data yielded by forced-choice assessments can be either ipsative, quasi-ipsative, or normative (Salgado et al., 2015). Ipsative data produced by a forced-choice measure refers to a measurement scale within a single person, not comparable to other people (Cattell, 1944). Clemans (1966), Hicks (1970), and Gleser (1972) documented the qualities and limitations of this kind of data type, such as constraints on the average correlation among scales ($\bar{r}$ is bounded by $-1/(d-1)$ and $(d-4)/d$, where d is the number of dimensions) and average correlation of scales with outcome ($\bar{r}_{xy} = 0$), and its consequent unsuitability for factor analysis and internal consistency reliability computation. Normative data refer to a measurement comparable across people, as produced by a Likert scale (or right/wrong data from a cognitive test). Such data are amenable to factor analysis and internal consistency reliability measurement. Quasi-ipsative data refer to data produced by a forced-choice procedure, but with properties not as limited as ipsative data and closer to normative data-like qualities. In the literature there are several approaches for generating quasi-ipsative data. One is using forced-choice methods scoring only a subset of the dimensions. As d increases, the limitations of ipsative data are less restrictive. A variant is the use of forced-choice methods with phantom dimensions, where the phantom (unscored) dimension is represented by a set of items that are not related to the targeted (scored) dimensions (Horn, 1971; Salgado & Lado, 2018). Another approach is the IRT methods developed to overcome the limitations of ipsative data (Brown & Maydeu-Olivares, 2013; Stark et al., 2005).

Modeling, Scoring, and Scaling

UNIDIMENSIONAL APPROACH Unidimensional pairwise preference task data are normative data and can be analyzed by classical test theory and sum score approaches and by IRT models for binary and polytomous data (Rasch [von Davier, 2016], Rasch rating scale [Andrich, 2016], Rasch partial credit [Masters, 2016], two-parameter logistic, 2PL [Birnbaum, 1968], generalized partial credit [Muraki & Muraki, 2016], and the graded response models [Samejima, 2016]). Multidimensional items with unscored dimensions (phantom dimensions) can also be analyzed using unidimensional methods if targeted dimension statements are only paired with unscored dimension statements (Horn, 1971; Salgado & Lado, 2018). Such assessments yield binary or polytomous data that reflect preference (or degree of preference) for the target dimension only (i.e., not also for the unscored dimension). These methods are easy to implement, although not as efficient, perhaps, as multidimensional methods.

UNFOLDING APPROACH The unfolding approach proposed by Stark et al. (2005) involves two steps, a pretesting calibration step with Likert rating responses, followed by the test administration with pairs. The data from the Likert rating pretesting are calibrated based on the generalized graded unfolding model (GGUM; J. S. Roberts et al., 2000). Pairs are assembled so that the two statements are from different dimensions (mostly) and are matched on social desirability. Social desirability can be based simply on the Likert statement mean or principal component score after negative keyed items are reflected (Bäckströmm & Björklund, 2013; Kuncel & Tellegen, 2009) or from a separate data collection with “fake good” instructions (Pavlov et al., 2021). Test administration provides binary data on the pairs (i.e., selecting statement i over k or not). The binary choice data is analyzed with the multi-unidimensional pairwise preference (MUPP) model. Stark et al. (2005) argued that separating the Likert rating calibration step from the analysis of pairs has advantages in simplifying computation and in having a calibrated statement pool that can be used to create adaptive tests.

However, the Stark et al. (2005) method relies on the assumption that the GGUM calibrations are invariant across the Likert and forced-choice format, even though the two administrations would typically involve different samples in different contexts perhaps even years apart; and the Likert statement stands by itself, whereas the forced-choice statement has a context of other statements in the item block. Also, Stark et al.’s approach does not lend itself to item analysis based on the blocks. So, P. Lee et al. (2019) proposed GGUM-RANK to estimate statement parameters and person parameters directly from the test administration of multidimensional forced-choice blocks without pretesting. It uses a generalization of the GGUM model proposed by Hontangas et al. (2015) in which ordering (e.g., $A > B > C$) is seen as a sequence of steps, selecting A versus B and C, then selecting B versus C. This enables the use of

triads rather than pairs (as in Stark et al., 2005). Hontangas et al. (2015) also showed that ranking methods were superior to other approaches and demonstrated how true scores, expected a posteriori (EAP) estimates, and traditional scores could be obtained and compared. P. Lee et al. (2019) demonstrated their approach in both simulations and with empirical data using Markov chain Monte Carlo (MCMC) estimation with a Metropolis–Hastings within Gibbs (Tierney) algorithm. They seem to have found that the triad format was more efficient than pairs, that statement discrimination within blocks was critical for estimation accuracy, and, most important, that the method performed well in providing convergent and discriminant validity evidence with external measures.

Thurstonian IRT and Other Dominance Approaches

Brown and Maydeu-Olivares (2011) and Maydeu-Olivares and Brown (2010) proposed Thurstonian IRT as a dominance alternative to the unfolding approach, basing it on Thurstone's (1927) law of comparative judgment. The method does not require pretesting and can be applied to forced-choice or ranking data of any size blocks. Response data are converted to $n(n - 1)/2$ binary pairwise preferences for the n statements in a block and analyzed. Brown and Maydeu-Olivares (2012) provided a tutorial on how to estimate the model using a confirmatory factor analysis approach with Mplus (Muthén & Muthén, 2001).

The basic concept is that each statement i elicits a latent utility, t_i (i.e., a degree of *desiredness* or *endorsement-worthiness* to respondents), and statement i is preferred over statement j if $t_i \geq t_j$. Utilities are modeled as $t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i$, where μ_i is the latent utility mean, λ_i is the factor loading of item i on the latent trait η_a , and ε_i is a normally distributed error with mean 0 and variance ψ_i^2 . The latent response variable (i.e., difference in utilities) is modeled as $y_{ij}^* = t_i - t_j = \mu_i - \mu_j + \lambda_i \eta_a - \lambda_j \eta_b + \varepsilon_i - \varepsilon_j$, which involves two latent traits (η_a, η_b), location parameters ($\mu_i - \mu_j$), factor loadings ($\lambda_i - \lambda_j$), and uniqueness parameter ($\psi_i^2 + \psi_j^2$). All these are specified as constraints in fitting a confirmatory factor analysis model in Mplus. Because of the complex error structure and often large number of latent traits, Thurstonian IRT models are typically estimated using limited-information methods.

Variations have been proposed. Wang et al. (2017) argued for the advantages of a one-parameter Rasch ipsative model (the Thurstonian IRT model is a two-parameter $[\mu, \lambda]$ model): It yields a single utility value for each statement, allows for between- and within-individual differentiation, and does not rely on statement blocks requiring mixed location (positive, negative) statements as Thurstonian IRT models do. This is an important issue because it can be a limitation of Thurstonian IRT models for high-stakes use (Bürkner et al., 2019). Morillo et al.'s (2016; Hontangas et al., 2016) MUPP-2PL model uses the MUPP choice model, but the 2PL instead of GGUM for calibration, making it a dominance model. Bunji and Okada's (2019) D-diffusion Thurstonian IRT model incorporates

response time information in the forced-choice item responses to improve model fit and increase the outcomes predictions. Response time information can inform response models (van der Linden, 2007); in personality, the assumption is that fast selection indicates easier option choices, such as stating “agree” to a binary Likert question (Ranger, 2013) or selecting an option in a forced-choice application (Bunji & Okada, 2019).

Reliability

Estimating reliability for unidimensional forced-choice models is straightforward, whether using CTT (e.g., alpha) or IRT (marginal reliability) approaches. Estimating reliability for the Stark et al. (2005) model is not straightforward. Hontangas et al. (2015) used a simulation study with known true scores and correlated them with EAP and traditional scores to show that EAP scores were more reliable. However, they did not address how reliability could be computed without knowing true scores. Seybert and Becker (2019) computed test–retest reliabilities, finding them to be only moderate and slightly lower than Likert scale versions of items. However, with rating scale test–retest reliability approaches, what is being measured is a mix of trait stability and stability of construct-irrelevant factors such as response style. Forced-choice methods are intended to reduce the proportion of response style variance and therefore should have lower test–retest factor correlations.

Estimating reliability with multidimensional forced-choice measures is discussed by Brown and Croudace (2015). They introduce the marginal reliability concept which can be estimated as theoretical or empirical reliability. Theoretical reliability involves averaging the model-based squared standard error for all trait values in the theoretical distribution, which requires multidimensional integration. Empirical reliability is preferred with many dimensions. This involves averaging squared standard errors of estimated trait scores in the sample, depending on the estimator used. For EAP, this is the posterior likelihood variance; for Maximum Likelihood (ML), the Fisher information inverse; for maximum a posteriori (MAP), the posterior information inverse. Brown and Maydeu-Olivares (2018) provided the solution for computing reliability from matrices produced by Mplus and provide R code for computing item information, standard errors, and reliability in the online supplement to that article. Mplus beginning in version 8.1 provides standard errors for MAP scores for IRT models of any dimensionality, which can be used to compute empirical reliability.

Validity and Fairness

The purpose of forced-choice and ranking approaches is to minimize the effects of response distortion but without altering the construct score. There are two situations in which this is important: high-stakes assessment, in which socially desirable responding or faking occurs; and for comparing individuals or groups with different response styles.

Faking effects can be examined experimentally: A treated group is asked to “fake good” to increase prospects for getting the job. Score inflation, the difference between scores

for the control and fake good groups, is less with forced choice compared to Likert-style measures (Cao, 2016; Cao & Drasgow, 2019), particularly when statements within a block are balanced in social desirability and normative approaches (e.g., IRT as discussed here) used for scoring. Bartram's (2007) meta-analysis showed that correlations with external measures (manager competency ratings) were higher with a forced-choice ($r = .38$) compared to a rating scale format ($r = .25$), in a study where managers used both. (With $N = 1,460$ and $r = .50$ across formats, this is a significant difference.)

A meta-analysis by Salgado and Tauriz (2014) showed that the multidimensional forced-choice format had substantially higher correlations with job performance and academic outcomes than the unidimensional forced-choice format. For Conscientiousness, rho (correlation between the predictor and outcomes corrected for predictor and criterion reliability and indirect range restriction in the predictor) was .40 for quasi-ipsative measurement (which includes both IRT and non-IRT multidimensional forced-choice assessments) but only .16 for normative forced choice (i.e., unidimensional). And this value, .40, is larger than any other meta-analyses of these variables, which tend to be dominated by Likert-scale measures.

A caveat to these positive findings is that cognitive ability may moderate these relationships such that brighter individuals may be more able to fake (Christiansen et al., 2005). However, Kyllonen et al. (2020) demonstrated incremental prediction beyond the GMAT/GRE and undergraduate GPA in predicting graduate business school grades and leadership activity for a forced-choice assessment administered during the admissions process. The increment in adjusted R -square was approximately .04 to .07 (depending on analysis details).

Forced-choice methods have advantages for evaluating culture/language group differences. Bartram (2013) found that at the individual level forced-choice and rating scale measures agreed moderately, but at the country level, correlations with external economic variables (the United Nation's Human Development Index, the Global Competitive Index) were more sensible with forced-choice measurement: Forced-choice Conscientiousness showed positive correlations with the economic variables, while rating scale Conscientiousness showed negative correlations. Country-level correlations do not have to follow individual-level correlations (ecological fallacy), but the forced-choice finding is in line with expectation and the rating scale finding is puzzling. In PISA 2012, two forced-choice scales (subject matter preference and learning style preference) were administered in a rating scale format in 2003 (same statement content). The rating scale format led to high correlations between dimensions and near-zero correlations with achievement. The forced-choice format led to expected negative correlations with each other (as an ipsative measure), but math preference had a high correlation with math achievement, as one would expect.

The studies cited here support the conclusion that forced choice improves the quality of information elicited with a survey instrument, at least when steps are taken to address the problems created by ipsative measurement. Designing and scoring forced-choice assessments is challenging—these are complex models relying on many assumptions.

And there are many questions remaining on how to improve the methodology. There are also user experience challenges. Respondents often complain about having to choose between two equally attractive (or unattractive) choices, even if test-taker motivation is unaffected (Sass et al., 2020). Still, forced-choice measurement is a promising methodology that ought to be included routinely in survey assessments of noncognitive skills, either as a supplement to or as a replacement for rating scale measures, particularly for high-stakes and cross-cultural assessments.

Situational Judgment Tests

Situational judgment tests (SJTs; also practical intelligence measures; McDaniel & Whetzel, 2005) have been popular assessment methods for organizational recruitment and selection (U.S. Office of Personnel Management, 2018a). This may be because of their high face validity (compared to standardized tests), low administrative costs (compared to interviews), relatively moderate subgroup differences (compared to standardized tests; Oostrom et al., 2015), lower susceptibility to faking (compared to personality tests; Kasten et al., 2018), and incremental outcome prediction beyond cognitive ability and personality (Clevenger et al., 2001; McDaniel et al., 2007). SJTs are also being considered in education for undergraduate (Oswald et al., 2004), business school (Hedlund et al., 2006), medical school (Patterson et al., 2016), and dental school admissions (Buyse & Lievens, 2011). SJTs can in principle be used to measure various skills but are mostly designed to measure interpersonal skills (Christian et al., 2010). The American Association of Medical Colleges is pilot testing SJTs for medical school admissions to measure eight primarily interpersonal competencies including service orientation, social skills, cultural competence, teamwork, and ethical responsibility (Ellison et al., 2024). SJTs are often ideally suited for education, training, and development purposes as well. Cox et al. (2017) presented evidence for advantages of SJTs over traditional training.

There are many kinds of SJTs, varying by intended construct (Chan & Schmitt, 2005; Christian et al., 2010; McDaniel & Whetzel, 2005), length (Crook et al., 2011), textual versus video items (Chan & Schmitt, 1997), instructions (Ployhart & Ehrhart, 2003), and other features (see reviews by Campion et al., 2014; Lievens et al., 2008; Oostrom et al., 2015; and Whetzel & McDaniel, 2009). Here, we focus on the most important considerations.

Ratings Versus Rankings

SJT items comprise a stem (situation description) followed by a set of response options, typically four to eight, but sometimes as few as one (Crook et al., 2011). Respondents can either rate or rank options. The number of rating categories typically ranges from two (e.g., “effective” vs. “ineffective”) to seven or more (e.g., “not at all effective” to “very effective”). Rankings can be full (“rank all options from best to worst”), partial (“choose the best option and choose the worst option”), or single choice (“choose the best option”). Arthur et al. (2014) concluded that ratings were superior to rankings for SJTs based on their higher correlations with personality,

but personality measurement is ratings based and therefore susceptible to spurious associations with ratings-based SJTs. The disadvantages of ratings are discussed in “Self-Report Ratings” and the advantages of rankings over ratings are discussed in “Forced-Choice and Ranking Methods.”

Best-Choice (Knowledge, “Should Do”) Versus Self-Prediction (Behavioral Tendency, “Would Do”) Instructions

Best-choice instructions ask respondents to choose, rank, or rate each option for its effectiveness. Self-prediction instructions ask respondents to choose, rank, or rate options by how they match to what they would do. Best-choice instructions lead to comparatively higher correlations with cognitive ability. Self-prediction instructions lead to comparatively higher correlations with personality (McDaniel et al., 2007). Best-choice instructions are less susceptible to faking than self-prediction instructions (Broadfoot, 2006; Nguyen et al., 2005; Peeters & Lievens, 2005), although responses to what others would do may be even less susceptible to faking (Lievens et al., 2009; Oostrom et al., 2017). Because of their resistance to faking and because knowledge instructions seem to be conceptually more straightforward than self-prediction instructions for eliciting responses, for some applications best-choice instructions may be better suited for SJTs than behavioral tendency instructions from a validity interpretative argument perspective.

Dimensionality and the Constructs Measured by SJTs

SJTs are typically found to be empirically unidimensional, even when constructed to reflect multiple dimensions. Oswald et al. (2004) wrote SJT items reflecting 12 dimensions of college student performance (e.g., leadership, artistic, career, perseverance, ethics), but found a strong general factor accounting for three times the variance of a second factor and only uninterpretable lower order factors. This is common for many constructs in which tests are developed to a framework that specifies several dimensions but a strong general factor accounts for most of the variance among items (Haberman et al., 2009). PISA 2012 specified a number of mathematical content (change and relationships, space and shape, quantity, uncertainty and data) and process (formulating situations, employing concepts, interpreting outcomes) dimensions (OECD, 2013, pp. 28–36), but found little empirical support for their differentiation, instead finding very high latent correlations among process types (all $r \geq .96$) and among content areas (all $r \geq .84$) (OECD, 2014, Table 12.9, p. 231). In such cases, dimensions serve to ensure construct coverage even though empirically they do not identify separate factors. The persistent empirical finding of unidimensionality has prompted discussions of broad factors underlying SJTs. Christian et al. (2010) reported that SJT developers targeted leadership (38% of SJTs), interpersonal skills (13%), personality (10%), teamwork (4%), job knowledge (3%), and heterogeneous composites (33%). However, SJTs might merely reflect broad tacit knowledge and practical intelligence (Chan & Schmitt, 2005; Sternberg & Horvath, 1999) or general domain knowledge (Corstjens

et al., 2017). Responses may alternatively reflect implicit trait policies, beliefs about the effectiveness of responses to situations based on effective trait expression (choosing an agreeable response when the situation calls for Agreeableness) (Motowidlo & Beier, 2010). There have been attempts to address multidimensionality in SJTs through construct explorations (Guenole et al., 2017; Westring et al., 2009). Still, it seems empirically that unidimensionality is found in SJTs even when multidimensionality is intended.

Scoring

Scoring SJTs can be more complex than scoring typical standardized tests: There can be disagreement on the best response to a situation, providing no single correct answer, and response options may vary continuously in appropriateness. SJT scoring approaches have been reviewed focusing on either ranking (Bergman et al., 2006) or rating methods (Weng et al., 2018; Whelpley, 2014). Bergman et al. (2006) discussed empirical keying (based on an outcome measure), theoretical scoring (based on keying options to a theoretical framework), expert scoring (keying to expert judgments), factor scoring, and subgrouping (based on clustering response patterns). Ratings-based scoring methods tend to focus on determining a profile similarity match between a test taker's rating and aggregate expert profile (McDaniel et al., 2011). De Leng et al. (2017) compared 28 ratings-based (profile similarity) measures, varying the reference group (experts versus the test-taking sample), agreement (simple percentage versus consensus), consensus standardization approach (raw/no standardization versus within-person standardization versus rating scale dichotomization), distance metric (absolute versus squared profile distances), and reference group aggregation method (mean, median, mode). Raw consensus, which credits responses by the proportion of the reference group selecting that response (e.g., if 20% of the reference group selects "agree," then respondents who select "agree" are credited .2 score points), led to higher alphas, particularly when the sample was the reference group and their judgments were aggregated using the mean. But Weng et al. (2018) found that raw consensus scoring led to lower correlations with external measures and suggested that extreme responses were problematic and that the value of midrange items was low.

Problems in SJT scoring may partially be attributable to the inherent problems with rating scales. Forced-choice or ranking approaches mitigate these problems and may therefore be better suited for SJTs. Zu and Kyllonen (2020) compared several observed score (including consensus) and IRT scoring methods for scoring forced-choice/ranking SJT methods. They found that consensus methods (using either experts or the sample as the reference group) were the best observed score approaches, but the best overall approach was the NRM with the sample as reference group. Guo et al. (2019) showed that although the NRM is a data-driven method, it produced an invariant scoring key across cohorts. Guo et al. (2016) demonstrated the value of a nonparametric item analysis approach for understanding SJTs, because SJTs often do not have a clear right or wrong answer, and of NRM for scoring.

Reliability

A well-known limitation to SJTs is their low internal consistency reliability (test–retest reliability is higher; Neubauer & Hofer, 2022). Kasten and Freund's (2016) meta-analytic reliability generalization study of SJTs in 271 studies found an average reliability of $\alpha = .61$ (standard deviation, $s = .20$). Factors depressing reliability were high-stakes use, item complexity, high fidelity, fewer items, pick-only response format, and earlier publication dates (i.e., the community is getting better at writing SJTs), but there were no effects for instruction type (knowledge versus behavioral tendency). These findings support the idea that SJTs are typically too short for operational use and that strategies for extracting more information (e.g., pick best and worst) are useful.

Rating measures tend to be more reliable than ranking methods. But reliability of rating methods confounds construct-irrelevant response style variance with construct-relevant trait variance, whereas ranking methods do not. Therefore, this apparent rating advantage cannot be interpreted as providing better measurement.

Validity

There is a strong argument for the validity evidence based on content supporting inferences drawn from SJTs. The development methodology typically involves a careful analysis of the performance situation SJTs are intended to reflect or predict, using critical incidents (see the section on behaviorally anchored rating scales, which uses the same analysis) or related methods to define the intended construct (Sternberg et al., 2000). The method itself asks respondents either what they would do or what the best response to the situation would be; either question should elicit useful respondent information allowing for a straightforward interpretation. The option ratings version of SJTs has the same conceptual problems as rating scales have generally (see the section on rating scales), and therefore the option ranking approaches have interpretive advantages from a validity standpoint.

There are several challenges to the interpretation of SJT responses in a straightforward manner. One is that SJTs ask respondents what they would or should do, but both of those are distinct from an actual response in a real-world situation. Ployhart and Ehrhart (2003) found relatively low correlations between responses in would-do and should-do formats. However, at least with best-choice (knowledge) instructions it could be argued that knowledge is a prerequisite to behavior. With behavioral tendency (would-do) instructions the matter is more muddled because some participants might respond as if to knowledge instructions (should-do) for fear of admitting to poor judgment.

Another challenge relates to a long-standing debate in the literature on what SJTs measure. Oswald et al. (2004) developed SJT items based on 12 college performance dimensions (e.g., leadership, artistic). Interrater agreement on item dimension assignment supported the distinctiveness of 12 dimensions (items with less than 75% agreement were removed), but factor analysis of response data did not, resulting in the adoption of a single score from the measure (Schmitt et al., 2009). Evidence for a general SJT factor is a common occurrence. D. J. R. Jackson et al. (2016) conducted a generalizability theory analysis on three different SJTs and over 10,000 job candidates

finding candidate (person) main effects, with little variation due to dimensions and situations, suggesting a general judgment factor across job situations. For many SJTs the situation descriptions themselves may not be central to what construct is being measured (Schaepers et al., 2020).

A general judgment factor is not the same as the general factor in cognitive tests. SJTs typically add to the prediction given by cognitive tests, as well as to the prediction given by personality measures. McDaniel et al.'s (2001) meta-analysis found a correlation of $\rho = .34$ between SJTs and job performance and $\rho = .46$ between SJTs and cognitive ability; SJTs add incrementally over cognitive ability in predicting outcomes (O'Connell et al., 2007). In education, several studies (Hedlund et al., 2006; Oswald et al., 2004; Schmitt et al., 2009) found that the general judgment factor added to other variables in predicting college outcomes, indicating SJTs are capturing something other cognitive variables are not.

Fairness

A consistent SJT finding is smaller subgroup differences compared to cognitive tests (Arthur et al., 2014; Clevenger et al., 2001; Oostrom et al., 2015; Oswald et al., 2004). There have been few measurement invariance studies of situational judgment, but Prasad (2017) found evidence that interests and socioeconomic status might be related to differential item functioning. For SJTs especially, one might expect culture differences. However, Prasad et al. (2017) found evidence for differences between Chinese versus U.S. respondents on some scales from a college performance SJT, but no differences overall.

Performance Measures of Noncognitive Skills

The assessments reviewed thus far can be called *descriptive assessments*, ones based on how respondents describe themselves (or how others describe them) or what they say they would or should do in a situation. Descriptive assessment approaches are different from cognitive tests in which respondents are asked to perform—to solve problems, recall information, or make a decision—which can be called *performance tests*. There are descriptive cognitive assessments. Mathematics self-efficacy items (“How confident do you feel about having to do the following mathematics tasks? ‘solving an equation like $3x + 5 = 17$.’ Very confident, Confident, Not very confident, Not at all confident”) and mathematics self-concept items (“I learn mathematics quickly. Strongly disagree, disagree, agree, strongly agree”; PISA 2012) are descriptive cognitive assessments. Data from performance tests compared to descriptive assessments undoubtedly inspire more confidence from stakeholders in assertions about student proficiency. This difference between descriptive assessments and performance tests might even account for the slow acceptance of noncognitive assessments despite demand for noncognitive skills. An important question is whether it is possible to make a performance-based noncognitive test.

Performance measures of noncognitive attributes have a long tradition; Cattell and Warburton (1967) referred to them as T-data (test data) measures. French (1948) presented students with a difficult number-series test with instructions stating that some problems had no solutions and that they could receive full credit by marking them as such. It was a disguised persistence test. French found that persistence scores were unrelated to other cognitive tests but were correlated with grades. A modern version is Alan et al.'s (2019) grit game. It comprises two versions of a number grid task in which the goal is to find three pairs of numbers that add up to 100. An easy version has 10 numbers, 5 of which are multiples of 10 and 2 of which are single digit (almost all students successfully completed this task). The difficult version has 20 double-digit numbers and only 1 is a multiple of 10 or 5 (the success rate was much lower). The reward for solving the problem within a short time limit is one gift for the easy version and four gifts for the hard version. Following a treatment designed to improve noncognitive skills by changing students' beliefs about the importance of effort, treated students were found to be more likely to choose the difficult version and 2 years later experienced a $d = .2$ effect on a standardized mathematics achievement test.

Segal's (2012) Coding Speed test, a word-number-lookup-matching task, was administered under low-stakes conditions to 12- to 16-year-old National Longitudinal Study of Youth participants. Segal interpreted the score from this test as a measure of intrinsic motivation because (a) an experiment found that performance improved for a subset of participants when given a financial (\$10) incentive, whereas others (the intrinsically motivated) performed at the same level regardless; and (b) test takers taking the test under high-stakes condition outperformed a more highly educated group taking it under low-stakes conditions. Segal found that performance on this test in the low-stakes National Longitudinal Study of Youth condition, controlling for cognitive ability, predicted earnings 23 years later, such that a 1 standard deviation increase in Coding Speed test performance was associated with a 9.5% increase in earnings.

In behavioral economics, *real-effort* tasks, such as Alan et al.'s (2019), are ones on which the respondent works, and the outcome depends on that work. This contrasts with *stated-effort* tasks in which respondents indicate the amount of effort or actions they would take in a situation (Charness et al., 2018). This distinction is analogous to the revealed versus stated preference distinction (Schl pfer & Fischhoff, 2010) and to the distinction we draw here between descriptive assessments and performance tests.

Real-effort task examples include solving mazes and anagrams, adding two-digit numbers, counting zeros in a grid, transcribing meaningless symbols, cracking walnuts, data entry, classifying Amazon reviews, packing quarters into boxes, playing a labyrinth game, digit span, filling envelopes, dragging a ball on a screen, typing alternating keys, and repeatedly typing the same paragraph, among others (Charness et al., 2018). Some of these also measure other factors (cognitive ability, physical skills). However, many could be used as performance tests of factors such as persistence.

Kyllonen and Kell (2018) reviewed tasks that can be understood as performance measures of noncognitive skills. They suggested several categories: cognitive test performance under low-stakes conditions (Segal's 2012 Coding Speed); objective personality tests in the Cattell–Warburton tradition (Kubinger, 2009; Ortner & Proyer, 2015; Ortner & Schmitt, 2014); economic preference tasks, such as time, risk, and social preferences (A. Falk et al., 2015, 2016); and confidence judgments for their calibration to actual performance (Stankov et al., 2013). They also reviewed measures based on test and survey behavior, such as survey effort as indicated by skipping or careless answering on background questions (Hitt et al., 2016; Zamarro et al., 2018), item position effects reflecting less effort on later items (Debeer et al., 2014; Weirich et al., 2017), and hasty response time reflecting lack of effort (Y.-H. Lee & Jia, 2014; Wise & Gao, 2014).

Several other lines of research evaluate performance measures of noncognitive skills. Collaborative problem solving (Hao et al., 2017) combines cognitive and interpersonal skills (OECD, 2017b). A current challenge is assigning credit to individual members (Hao et al., 2019; Martin-Raugh et al., 2020). Idea generation (Bennett & Rock, 1995) can be considered a standard cognitive task, but there are additional dispositional aspects in choosing to respond with more than one solution or more than one answer to a question. Critical thinking also involves dispositional aspects.

Two additional performance measures of noncognitive skills, attitudes, or worldviews have been researched. Implicit association tests (Greenwald et al., 1998) involve interpreting patterns of response times on classification tasks as indicators of underlying attitudes and beliefs. Although mainly known as an implicit bias measure, it can be used to measure personality (Grumm & von Collani, 2007). Conditional reasoning (James, 1998; LeBreton et al., 2020), measures a personality disposition or worldview through multiple-choice answer selection to a reading passage with two correct answers. One answer is consonant with one belief system or worldview and a different answer with a different belief system.

Reliability, Validity, and Fairness

Performance measures are diverse, making summary statements about psychometric issues difficult. However, tasks discussed can be treated with standard psychometric procedures (e.g., IRT; some have, Debeer et al., 2014). Also, fairness concerns are typically ignored. As the use of performance measures increases, we are likely to see increased attention paid to these issues, following guidelines such as those captured in the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Challenges associated with these measures include alternative interpretations. Many real-effort tasks likely reflect both the target construct, a personality factor, and construct-irrelevant cognitive or physical ability factors. Watts et al.'s (2018) review of the marshmallow test, an indicator of a child's ability to delay gratification, found that much of its association with later education and work achievements was attributable to sociodemographic and cognitive ability factors. Controlling for

construct-irrelevant influences may be especially important for performance tasks. Construct underrepresentation is another challenge. Skipping an item on a background questionnaire may be a weak indicator of Conscientiousness, but there are other elements of Conscientiousness not captured with such a measure. It will likely prove fruitful to combine multiple and diverse such indicators to create composite construct measures.

Biodata, Documented Accomplishments, and Administrative Records

Choices made and experiences had influence life directions. Choices and experiences might be gathered from an informant who knows the target, captured with a questionnaire (or personal statement), or reflected on one's resume or bio. Life record data (L-data) was one of three data types identified by Cattell (1965), with T-data and Q-data (questionnaire data). L-data is defined as "behaviour in the actual, everyday life situation" such as "number of automobile accidents over 20 years,' 'frequency of engagements,' 'number of societies to which the person belongs,' 'marks in school,' and so on" (Cattell, 1965, p. 61). Cattell suggested that someone who knows the person well might be able to provide such data.

Organizational psychology has a long history of research on biodata (biographical data), responses to survey questions about one's own background and experiences. Oswald et al. (2004) included a biodata measure along with the SJT to reflect 12 dimensions of college experience (see the section on SJTs). Some biodata questions were, "How often have you signed a petition for something you believed in? very often/often/sometimes/seldom/never" (citizenship) and "How many times in the last year have you tried to get someone to join an activity in which you were involved or leading? Never/once/twice/three or four times/five times or more" (leadership). Stricker and Rock (1998) reported on an accomplishments questionnaire (designed for graduate school applicants) with items such as "conducted a band, orchestra, or vocal group at a public performance"; "wrote poetry, fiction, or essays that were published"; and "was elected to a major class office in college." These were drawn from dimensions of academic achievement, leadership, language, aesthetics expression, science, and mechanical. Mumford and Owens (1987) and Mumford et al. (2012) reviewed the history, measures, and methodology of biodata, referred to as background data. Biodata measures have been found to be comparable in their predictions of outcomes to Conscientiousness ($r = .30$), although they overlap considerably with cognitive ability, $r = .50$ (Schmidt & Hunter, 1998). Related are behavioral measures of personality (Breil et al., 2022; J. J. Jackson et al., 2010; Soto et al., 2022) based on the specific behaviors individuals engage in, and ambulatory assessments (Trull & Ebner-Priemer, 2013), such as diaries and event sampling.

Also related are work-sample tests for college admissions. The measure is not what one claims on a survey but a sample of actual behavior. Niessen et al. (2016; Meijer & Niessen, 2015; Niessen & Meijer, 2017) provided several suggestions for how such a

behavioral-sampling system could be implemented, such as trial studying approaches (applicant attends a lecture, studies material, takes an exam), and multiple mini-interviews (similar to SJTs) in which applicants are presented with problems they might encounter in professional practice and asked how they would approach them.

As organizations and schools continue to develop increasingly sophisticated and comprehensive data storage systems, such as New Jersey's longitudinal student data warehouse (NJ Smart) or the U.S. National Student Clearinghouse (transcripts and enrollment data), biodata studies can be conducted solely with administrative records without burdening students or teachers. Kautz and Zantoni (2024) combined grades, accumulated credits, absences, and disciplinary infractions, all obtained from school records, to create a noncognitive composite that was as strong a predictor of a variety of future outcomes (11th-grade grades, absences, arrests, graduation, college enrollment, college graduation) as cognitive ability scores were. K. Jackson (2018) combined absences, suspensions, course grades, and grade repetition as a supplement to achievement test scores to create a teacher value-added measure that was more predictive of lagged student outcomes than were test scores alone, which is the typical value-added measure. (Value-added modeling is a controversial approach in educational accountability; see American Statistical Association [2014] and Ho and Polikoff (this volume); the point here is to show that including noncognitive information can potentially improve such measurement. See also C. K. Jackson et al., 2021.)

With advances in data collection technology and statistical methods, big data and machine learning approaches (e.g., regularization techniques to handle wide data [more variables than people]; gradient boosting machines for better out-of-sample prediction), and psychometrics (e.g., latent classes and mixture models to replace biodata's subgrouping methods), there is an opportunity to make advances in biodata measurement. In the early 21st century, such data are much easier to gather through administrative records retained routinely in database systems or on the web. Time spent studying or working on homework is an important determinant of learning (Cooper et al., 2006; Grodner & Rupp, 2013) and is now easier to capture with online learning systems. Variables like registration latency as indicators of procrastination can be captured (Novarese & Di Giovinazzo, 2013). Social media activity relates to personality factors (Kosinski et al., 2013; Youyou et al., 2015), as do contents of offices and workspaces (Gosling et al., 2002). Mobile sensing data from phones, wearables, and beacons differentiate workplace performance levels (Mirjafari et al., 2019). Stress levels can be detected through speech (Slavich et al., 2019). There likely will be rapid developments in these areas.

CONCLUSION AND PROSPECTS

This chapter reviewed noncognitive constructs and methods for assessing them. Noncognitive skills are important determinants and outcomes of education and are in high demand in the workplace. An education benefit is noncognitive in addition to

cognitive skills development, but this benefit is not always apparent because we do not routinely monitor noncognitive skill development. Instead, we infer this relationship because the positive benefits of schooling on a variety of workforce and life outcomes greatly exceeds the benefit attributable to the development of cognitive skills alone. We also know from research involving the administering of noncognitive measures, almost exclusively self- or others reports collected with rating scales, that such measures predict a wide variety of education, workforce, and life outcomes. This relationship is sometimes as strong as and sometimes stronger than the relationship between cognitive measures and those outcomes. However, a recurring finding is that the relationship between noncognitive skills and outcomes is moderated by cognitive ability—the relationship is particularly strong for students of lower cognitive ability and workers. Analyses of the nature of jobs in the workforce, particularly with technology-induced change, suggest that noncognitive skills are likely to increase in importance.

A roadblock to adoption of noncognitive measurement in schools is a widespread but misguided perception that personality is an immutable trait outside the influence of a teacher or the educational system. But ample research demonstrates that personal qualities such as work ethic, social skill, and productive attitudes are as amenable to change due to good teaching and a supportive climate as curricular achievement is. Another barrier has been simple terminology—throughout this article we have used a variety of terms such as personality, social-emotional learning, soft skills, noncognitive skills, interpersonal and intrapersonal skills, attitudes, values, and beliefs almost interchangeably. Each term has a distinct history and constituency but there is a growing recognition that all these terms refer to skills that are fundamentally in the same realm as cognitive skills are. What has largely distinguished noncognitive skills is not the nature of the constructs per se as much as it is the measurement method—cognitive skills tend to be measured with tests and noncognitive skills with ratings. Since 2015 there has been an expansion of frameworks, accompanied by compendia of measures, outlining the nature and extent of noncognitive skills, and we now have a good understanding of what some of the most important skills are.

The literature and practice of noncognitive assessment has been almost completely devoted to the use of rating scale measures, primarily self-ratings. This is due to their ease of development, administration, scoring, and reporting, but there are serious limitations to rating scale measures. They are subject to various unmonitored and uncontrolled biases (reference, response style, social desirability) and the adoption of self-reports limits their use to low-stakes applications. Alternative approaches have benefits—others' reports, forced-choice and ranking approaches, anchoring methods, SJTs, various performance tests of noncognitive skills such as real-effort measures of Conscientiousness, as well as biodata and administrative records, and big data (data mining) approaches. Any measure has advantages and disadvantages in certain situations and it is important to consider the relative strengths and weaknesses of alternatives to determine their suitability for particular applications. Reliability, validity, and fairness analyses can help. Modern psychometric methods—CTT, generalizability theory, IRT (see Cai et al., this volume,

for a discussion of models)—are suitable for almost all noncognitive measurement, even though adoption of proper psychometric methods is not yet widespread. An interpretive-argument perspective on validity is an appropriate filter through which to evaluate noncognitive measurement methods, the data they produce, and the inferences we wish to draw from those data. A major distinction is between low-stakes use, as is currently widely implemented in schools, and high-stakes use, which is growing in workforce personnel selection and increasingly in higher education. Likert-scale self-measurement is inadequate for high-stakes use, but alternatives, such as others' ratings, forced-choice methods, and SJTs, may be adequate for high-stakes use, particularly when treated with appropriate psychometric methods and after pilot testing for susceptibility to coaching and faking. It may also be important to use multiple measurement methods and indicators to mitigate the weaknesses of any one method.

A general theme throughout this chapter has been that the level of interest and value attributed to noncognitive skills in both education and workplace settings, from both the practice and the policy communities, has outpaced the field's ability to provide adequate measures of those skills. We have evidence that interest and perceived importance are likely to grow. It is essential that the educational measurement community meets that interest with appropriate tools.

ACKNOWLEDGMENTS

Preparation of this chapter was supported by funding from the ETS Research Institute, National Science Foundation Award 2201888, and National Institutes of Health Award R01 HD107079. We thank Ida Lawrence, the editors Linda Cook and Mary Pitoniak, and reviewers Fred Oswald, Rob Meijer, and Suzanne Lane, for their inspiring, supportive, constructive, and often extensive comments, which improved the manuscript immeasurably.

REFERENCES

- Aicheson, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139–160. <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3), 1121–1162. <https://doi.org/10.1093/qje/qjz006>
- Almlund, M., Duckworth, A., Heckman, J. J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 4, pp. 1–181). Elsevier.
- Alreck, P. L., & Settle, R. B. (1994). *The survey research handbook, second edition: Guidelines and strategies for conducting a survey*. McGraw–Hill.

- American Association of Medical Colleges. (2020). *AAMC situational judgment test pilot*. <https://students-residents.aamc.org/applying-medical-school/preparing-med-school/aamc-situational-judgment-test-pilot/>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- American Statistical Association. (2014, April 8). *ASA statement on using value-added models for educational assessment*. <https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf>
- Andrich, D. (2016). Rasch rating-scale model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 75–94). Routledge.
- Arthur, W., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology*, 99(3), 535–545. <https://doi.org/10.1037/a0035788>
- Aspen Institute. (2018). *The practice base for how we learn: Supporting students' social, emotional, and academic development. Consensus Statements of Practice from the Council of Distinguished Educators* (Sheldon Berman with Sydney Chaffee & Julia Sarmiento). https://assets.aspeninstitute.org/content/uploads/2018/03/CDE-Practice-Base_FINAL.pdf
- Asplund, J., Agrawal, S., Hodges, T., Harter, J., & Lopez, S. J. (2014). *The Clifton StrengthsFinder® 2.0 technical report: Development and validation*. The Gallup Organization.
- Assessment Work Group. (2019). *Student social and emotional competence assessment: The current state of the field and a vision for its future*. Collaborative for Academic, Social, and Emotional Learning.
- Atwell, M. N., & Bridgeland, J. M. (2019). *Ready to lead: A 2019 update of principals' perspectives on how social and emotional learning can prepare children and transform schools. A report for CASEL*. Civic with Hart Research Associates.
- Autor, D. H., Levy, F., & Murnane, R. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333.
- Autor, D. H., & Price, B. (2013). *The changing task composition of the US labor market: An update of Autor, Levy, and Murnane (2003)*. MIT Economics. <https://economics.mit.edu/people/faculty/david-h-autor/working-papers>
- Bäckströmm, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, 54(2), 152–159.
- Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017) Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 10(1), 7–39. <https://doi.org/10.1080/19345747.2016.1232459>
- Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, 41, 586–598.

- Barnett, S. (1996). *Lives in the balance: Age-27 benefit–cost analysis of the high/scope Perry preschool program* (High/Scope Educational Research Foundation Monograph No. 11). High Scope Press.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
- Barrow, L., & Rouse, C. E. (2005). Do returns to schooling differ by race and ethnicity? *American Economic Review*, 95(2), 83–87. <https://doi.org/10.1257/000282805774670130>
- Barton, D., Farrell, D., & Mourshed, M. (2012). *Education to employment: Designing a system that works*. McKinsey Center for Government. <https://www.mckinsey.com/industries/social-sector/our-insights/education-to-employment-designing-a-system-that-works>
- Bartram, D. (2005). The Great Eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, 90, 1185–1203.
- Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment*, 15(3), 263–272.
- Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology*, 44(1), 61–83. <https://doi.org/10.1177/0022022111430258>
- Baxter, J. C., Brock, B., Hill, P. C., & Rozelle, R. M. (1981). Letters of recommendation: A question of value. *Journal of Applied Psychology*, 66(3), 296–301. <https://doi.org/10.1037/0021-9010.66.3.296>
- Becker, G. S. (1994). *Human capital: A theoretical and empirical analysis with special reference to education* (3rd ed.). University of Chicago Press.
- Bennett, R. E., & Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test *Journal of Educational Measurement*, 32(1), 19–36. <https://www.jstor.org/stable/1435190>.
- Berg, J., Osher, D., Same, M. R., Nolan, E., Benson, D., & Jacobs, N. (2017). *Identifying, defining, and measuring social and emotional competencies—final report*. American Institutes for Research. <https://www.air.org/resource/identifying-defining-and-measuring-social-and-emotional-competencies>
- Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, 14(3), 223–235.
- Bertling, J. P. (2015, May 21–24). *Noncognitive modules for the National Assessment of Educational Progress* [Paper presentation]. Association for Psychological Science 27th Annual Convention, New York City, NY, United States.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Addison–Wesley.

- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain*. Longmans, Green and Co.
- Blyth, D. A., Borowski, T., Farrington, C. A., Kyllonen, P., & Weissberg, R. P. (2019). *Ten criteria for describing and selecting SEL frameworks*. Collaborative for Academic, Social, and Emotional Learning.
- Blyth, D. A., Jones, S., & Borowski, T. (2018). *SEL frameworks—what are they and why are they important?* CASEL. <https://measuringsel.casel.org/wp-content/uploads/2018/09/Frameworks-A.1.pdf>
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51. <https://doi.org/10.1007/BF02291411>
- Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, 9(4), 453–465. <https://doi.org/10.1037/1082-989X.9.4.453>
- Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. *Behavior Research Methods*, 44, 144–157.
- Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2011). Identification problems in personality psychology. *Personality and Individual Differences*, 51(3), 315–320.
- Borghans, L., Golsteyn, B. H. H., Heckman, J. J., & Humphries, J. E. (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47), 13354–13359.
- Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, 64, 410–421.
- Bowles S., Gintis H., & Osborne M., (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 39(4), 1137–1176. <https://doi.org/10.1257/jel.39.4.1137>
- Bravo-Sanzana, M. V., Varela, J., Terán-Mendoza, O., & Rodriguez-Rivas, M. E. (2023). Measuring school social climate in Latin America: The need for multidimensional and multi-informant tests—A systematic review. *Frontiers in Psychology*, (14). <https://doi.org/10.3389/fpsyg.2023.1190432>
- Breil, S. M., Mielke, I., Ahrens, H., Geldmacher, T., Sensmeier, J., Marschall, B., & Back, M. D. (2022). Predicting actual social skill expression from personality and skill self-concepts. *Journal of Intelligence*, 10(3), 48. <http://dx.doi.org/10.3390/jintelligence10030048>
- Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.
- Broadfoot, A. A. (2006). *Response instructions and faking on situational judgment tests*. Bowling Green State University. http://rave.ohiolink.edu/etdc/view?acc_num=bgsu1161283237
- Brown, A. (2016) Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research*, 51(2–3), 345–356. <https://doi.org/10.1080/00273171.2016.1150152>

- Brown, A., & Bartram, D. (2009). *The occupational personality questionnaire revolution: Applying item response theory to questionnaire design and scoring* (Technical report). SHL Group. <https://kar.kent.ac.uk/74438/>
- Brown, A., & Croudace, T. J. (2015). Scoring and estimating score precision using multidimensional IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Multivariate applications series. Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307–333). Routledge.
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71(3), 460–502.
- Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*, 44(4), 1135–1147. <https://doi.org/10.3758/s13428-012-0217-x>
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18(1), 36–52. <https://doi.org/10.1037/a0030641>
- Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 516–529. <https://doi.org/10.1080/10705511.2017.1392247>
- Bunji, K., & Okada, K. (2019). Item response and response time model for personality assessment via linear ballistic accumulation. *Japanese Journal of Statistics and Data Science*, 2, 263–297. <https://doi.org/10.1007/s42081-019-00040-4>
- Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement*, 79(5), 827–854.
- Burrus, J., Jackson, T., Xi, N., & Steinberg, J. (2013). *Identifying the most important 21st century workforce competencies: An analysis of the Occupational Information Network (O*NET)* (ETS RR-13-21). ETS.
- Buyse, T., & Lievens, F. (2011). Situational judgment tests as a new tool for dental school selection. *Journal of Dental Education*, 75(6), 743–749.
- California Department of Education. (2018). *Social and emotional learning in California: A guide to resources*. <https://www.cde.ca.gov/eo/in/documents/selresourcesguide.pdf>
- Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Framework for enhancing education and workplace success* (ACT Research Report Series 2015 (4)). ACT. <https://files.eric.ed.gov/fulltext/ED558040.pdf>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. <https://doi.org/10.1037/h0046016>
- Campion, M. C., Ployhart, R. E., & MacKenzie, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance*, 27, 283–310.
- Cao, M. (2016). *Examining the fakability of forced-choice individual differences measures* [Doctoral dissertation]. University of Illinois. <http://hdl.handle.net/2142/93064>

- Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. <https://doi.org/10.1037/apl0000414>
- Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics*, Volume 3, Part A (pp. 1801–1863). Elsevier Science.
- Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2020). *The twenty-first mental measurements yearbook*. Buros Center for Testing.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
- Casner-Lotto, J., & Barrington, L. (2006). *Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the 21st century U.S. workforce*. Partnership for 21st Century Skills.
- Castex, G., & Dechter, E. K. (2014). The changing roles of education and ability in wage determination. *Journal of Labor Economics, 32*(4), 685–710.
- Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303. <https://doi.org/10.1037/h0057299>
- Cattell, R. B. (1965). *The scientific analysis of personality*. Penguin Books.
- Cattell, R. B., & Warburton, F. W. (1967). *Objective personality and motivation tests*. University of Illinois Press.
- Cengage. (2019, January 16). *New survey: Demand for "uniquely human skills" increases even as technology and automation replace some jobs*. <https://news.cengage.com/upskilling/new-survey-demand-for-uniquely-human-skills-increases-even-as-technology-and-automation-replace-some-jobs/>
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*(1), 143–159. <https://doi.org/10.1037/0021-9010.82.1.143>
- Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson, & O. Smit-Voskuil (Eds.), *The Blackwell handbook of personnel selection* (pp. 219–242). Blackwell Publishing.
- Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior and Organization, 149*, 74–87. <https://doi.org/10.1016/j.jebo.2018.02.024>
- Chernyshenko, O. S. (2003). Applications of ideal point approaches to scale construction and scoring in personality measurement: The development of a six-faceted measure of conscientiousness. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 63*(11-B), 5556.
- Chernyshenko, O., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and well-being: Conceptual framework for the OECD study on social and emotional skills* (OECD Education Working Papers, No. 173). OECD Publishing. <https://doi.org/10.1787/db1d8e59-en>

- Christian, M. S., Edwards, J. C., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, 63, 83–117.
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, 18(3), 267–307.
- Clark, D., & Martorell, P. (2014). The signaling value of a high school diploma. *Journal of Political Economy*, 122(2), 282–318.
- Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, 14, 1–56.
- Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology*, 86(3), 410–417.
- Cochrane, R. C. (1966). *Measures for progress: A history of the National Bureau of Standards*. U.S. Government Printing Office.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159.
- Collaborative for Academic, Social, and Emotional Learning. (2020a). *State scan scorecard project*. <https://casel.org/state-scan-scorecard-project-2/#info>
- Collaborative for Academic, Social, and Emotional Learning. (2020b). *SEL assessment guide*. <https://measuringSEL.casel.org/access-assessment-guide/>
- Condon, D. M. (2018). *The SAPA Personality Inventory: An empirically derived, hierarchically organized self-report personality assessment model*. PsyArXiv. <https://doi.org/10.17605/osf.io/SC4P9>
- Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322–328.
- Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136(6), 1092–1122.
- Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research*, 76(1), 1–62.
- Corstjens, J., Lievens, F., & Krumm, S. (2017). Situational judgment tests for selection. In H. W. Goldstein, E. D. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention*. John Wiley & Sons. <https://doi.org/10.1002/9781118972472.ch11>
- Costa, A. L., & Kallick, B. (2008). *Learning and leading with habits of mind: 16 essential characteristics for success*. Association for Supervision and Curriculum Development.

- Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Center.
- Cox, C. B., Barron, L. G., Davis, W., & de la Garza, B. (2017). Using situational judgment tests (SJTs) in training: Development and evaluation of a structured, low-fidelity scenario-based training method. *Personnel Review*, 46(1), 36–45. <https://doi.org/10.1108/PR-05-2015-0137>
- Credé, M., Tynan, M., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, 113, 492–511.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.
- Crook, A. E., Beier, M. E., Cox, C. B., Kell, H. J., Hanks, A. R., & Motowidlo, S. J. (2011). Measuring relationships between personality, knowledge, and performance using single-response situational judgment tests. *International Journal of Selection and Assessment*, 19(4), 363–373.
- Damian, R. I., Spengler, M., Sutut, A., & Roberts, B. W. (2019). Sixteen going on sixty-six: A longitudinal study of personality stability and change across 50 years. *Journal of Personality and Social Psychology*, 117(3), 674–695. <https://doi.org/10.1037/pspp0000210>
- Darr, W., Borman, W. C., St-Pierre, L., Kubisiak, C., & Grossman, M. (2017). An applied examination of the computerized adaptive rating scale for assessing performance. *International Journal of Selection and Assessment*, 25(2), 149–153.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.
- Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading assessment. *Journal of Educational and Behavioral Statistics*, 39(6), 502–523.
- Debnath, S. C., Lee, B. B., & Tandon, S. (2015). Fifty years and going strong: What makes behaviorally anchored rating scales so perennial as an appraisal method? *International Journal of Business and Social Science*, 6(2), 16–25.
- De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P., & Themmen, A. P. N. (2017) Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Science Education, Theory, and Practice*, 2, 243–265. <https://doi.org/10.1007/s10459-016-9720-7A>
- Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, 132(4), 1593–1640.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41, 417–440. <https://doi.org/10.1146/annurev.ps.41.020190.002221>
- District of Columbia Public Schools. (2017). *A capital commitment 2017–2022*. <https://dcps.dc.gov/capitalcommitment>

- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 3(4), 465–476.
- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support selection and classification decisions* (Technical Report 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Duckworth, A. (2013). *Grit: The power of passion and perseverance*. TED Talk Lesson. https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance
- Duckworth, A. (2016). *Grit: The power of passion and perseverance*. Scribner.
- Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, 44(4), 237–251.
- Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology*, 91, 40–57.
- Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory. *Educational and Psychological Measurement*, 79(1), 108–128. <https://doi.org/10.1177/0013164417752782>
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82, 405–432.
- Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology*, 45, 294–309.
- Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (Eds.). (2015). *Handbook of social and emotional learning: Research and practice* (pp. 3–19). The Guilford Press.
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.
- Dweck, C. (2014). *The power of believing that you can improve*. TEDxNorrköping. https://www.ted.com/talks/carol_dweck_the_power_of_believing_that_you_can_improve
- Dweck, C. (2017). *Mindset—updated edition: Changing the way you think to fulfill your potential*. Little, Brown Book Group.
- Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255–270). Springer.

- Ellison, H. B., Grabowski, C. J., Schmude, M., Costa, J. B., Naemi, B., Schmidt, M., Patel, D.; Westervelt, M. (2024). Evaluating a situational judgment test for use in medical school admissions: Two years of AAMC PREview Exam administration data. *Academic Medicine* 99(2), 183–191. 10.1097/ACM.00000000000005548
- Elvira, M., & Town, R. (2002). The effects of race and worker productivity on performance evaluations. *Industrial Relations*, 40(4), 571–590.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- ETS. (2014). *ETS standards for quality and fairness*. <https://www.ets.org/s/about/pdf/standards.pdf>
- ETS. (2022). *ETS guidelines for developing fair tests and communications*. <https://www.ets.org/pdfs/about/fair-tests-and-communications.pdf>
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2015). *The nature and predictive power of preferences: Global evidence*. Institute for the Study of Labor. <https://docs.iza.org/dp9504.pdf>
- Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2016). *The preference survey module: A validated instrument for measuring risk, time, and social preferences* (Discussion Paper No. 9674). Institute for the Study of Labor. <https://docs.iza.org/dp9674.pdf>
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 21(3), 328–347. <https://doi.org/10.1037/met0000059>
- Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment*, 93(5), 445–453.
- Farrington, C. A., Roderick, M., Allensworth, E. A., Nagaoka, J., Johnson, D. W., Keyes, T. S., & Beechum, N. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in academic performance—a critical literature review*. Consortium on Chicago School Research.
- Farruggia, S. P., Han, C. W., Watson, L., Moss, T. P., & Bottoms, B. L. (2016). Noncognitive factors and college student success. *Journal of College Student Retention: Research, Theory & Practice*, 20(3), 308–327.
- Feng, S., Han, Y., Heckman, J. J., & Kautz, T. (2022). Comparing the reliability and predictive power of child, teacher, and guardian reports of noncognitive skills. *Proceedings of the National Academy of Sciences*, 119(6), e2113992119. <https://doi.org/10.1073/pnas.2113992119>
- Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology*, 98(4), 659–667.
- Flake, J. (2017, August 18). Jeff Flake: We need immigrants with skills. But working hard is a skill. *New York Times*. <https://www.nytimes.com/2017/08/18/opinion/jeff-flake-we-need-immigrants-with-skills-but-working-hard-is-a-skill.html>
- Fowler, F. J. (2006). *Survey research methods* (3rd ed.). Sage Publications.

- Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39(3), 193–202. <https://doi.org/10.1037/0003-066X.39.3.193>
- French, J. W. (1948). The validity of a persistence test. *Psychometrika*, 13, 271–277. <https://doi.org/10.1007/BF02289223>
- Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, 138, 296–321.
- Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting and Social Change*, 114, 254–280.
- Fu, J., Tank, X., & Kyllonen, P. (2022). *Can the generalized graded unfolding model fit dominance statements?* [Unpublished manuscript]. ETS.
- Gensowski, M. (2018). Personality, IQ, and lifetime earnings. *Labour Economics*, 51, 170–183. <https://doi.org/10.1016/j.labeco.2017.12.004>
- Gleser, L. J. (1972). On bounds for the average correlation between subtest scores in ipsatively scored tests. *Educational and Psychological Measurement*, 32, 759–765.
- Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management*, 48(2), 251–268. <https://doi.org/10.1002/hrm.20278>
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26–34. <https://doi.org/10.1037/0003-066x.48.1.26>
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84–96.
- Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. Bantam.
- Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, 82(3), 379–398.
- Graham, S. (1991). A review of attribution theory in achievement contexts. *Educational Psychology Review*, 3, 5–39.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21(4), 347–360.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, 44(2), 93–109.
- Grumm, M., & von Collani, G. (2007). Measuring big-five personality dimensions with the implicit association test—implicit personality traits or self-esteem? *Personality and Individual Differences*, 43(8), 2205–2217. <https://doi.org/10.1016/j.paid.2007.06.032>
- Guenole, N., Chernyshenko, O. S., & Weekly, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International*

- Journal of Testing*, 17(3), 234–252. <https://doi.org/10.1080/15305058.2017.1297817>
- Guo, H., Zu, J., & Kyllonen, P. (2019). Consistency of NRM-based scoring rules across cohorts for a situational judgment test. *Psychological Test and Assessment Modeling*, 61, 207–225.
- Guo, H., Zu, J., Kyllonen, P., & Schmitt, N. (2016). *Evaluation of different scoring rules for a non-cognitive test in development* (ETS Research Report RR-16-03). ETS.
- Guttman, L., & Levy, S. (1982). On the definition and varieties of attitude and wellbeing. *Social Indicators Research*, 10, 159–174.
- Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics Press.
- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204–229. <https://doi.org/10.3102/1076998607302636>
- Haberman, S. J., Sinharay, S., & Puhane, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, 62, 79–95.
- Hall, E. P., Gott, S. P., & Pokorny, R. A. (1995). *A procedural guide to cognitive task analysis: The PARI methodology* (Technical Report ADA303654). Armstrong Lab, Brooks Air Force Base. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a303654.pdf>
- Hamilton, L. S., Stecher, B. M., Schweig, J., & Baker, G. (2018). *RAND Education Assessment Finder*. RAND Corporation. <https://www.rand.org/pubs/tools/TL308.html>
- Hao, J., Liu, L., von Davier, A. A., & Kyllonen, P. C. (2017). Initial steps towards a standardized assessment for CPS: Practical challenges and strategies. In A. A. von Davier, M. Zhu, & P. C. Kyllonen (Eds.), *Innovative assessment of collaboration* (pp. 135–156). Springer.
- Hao, J., Liu, L., Kyllonen, P., Flor, M., & von Davier, A. (2019). *Psychometric considerations and a general scoring strategy for assessments of collaborative problem solving*. [ETS Research Report No. RR-19-41]. ETS. <https://doi.org/10.1002/ets2.12276>.
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41(1), 43–62. <https://doi.org/10.1111/j.1744-6570.1988.tb00631.x>
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology*, 48(3), 319–334.
- He, J., & van de Vijver, F. J. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly*, 79, 267–290.
- Heckman, J. J., Humphries, J. E., & Kautz, T. (Eds.). (2014). *The myth of achievement tests: The GED and the role of character in American life*. University of Chicago Press.

- Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review*, 91(2), 145–149. <https://doi.org/10.1257/aer.91.2.145>
- Hedlund, J., Wilt, J. M., Nebel, K. R., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the Graduate Management Admissions Test. *Learning and Individual Differences*, 16, 101–127.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167–184.
- Hilton, M. (2019). Education for workforce readiness: Findings from reports of the National Academies of Sciences, Engineering, and Medicine. In F. L. Oswald, T. S. Behrens, & L. L. Foster (Eds.), *Workforce readiness and the future of work* (pp. 189–206). Routledge.
- Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review*, 52(C), 105–119.
- Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, 39(8), 598–612. <https://doi.org/10.1177/0146621615585851>
- Hontangas, P. M., Leenen, I., de la Torre, J., Ponsoda, V., Morillo, D., & Abad, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*, 28(1), 76–82.
- Horn, J. L. (1971). Motivation and dynamic calculus concepts from multivariate experiment. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (2nd printing, pp. 611–641). Rand McNally.
- Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement*. Policy Analysis for California Education. https://edpolicyinca.org/sites/default/files/SEL-CC_report.pdf
- Inchley, J., Currie, D., Cosma, A., & Samdal, O. (Eds.). (2018). *Health behaviour in school-aged children (HBSC) study protocol: Background, methodology and mandatory items for the 2017/18 survey*. CAHRU.
- Institute of Educational Sciences. (2007). *Character education: WWC evidence review protocol for character education interventions*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/CharEd_protocol.pdf <https://eric.ed.gov/?id=ED497054>
- Institute of Educational Sciences. (2020). *WWC What Works Clearinghouse*. U.S. Department of Education. <https://ies.ed.gov/ncee/wwc/>
- Ion, A., Gunnesch-Luca, G., Petre, D., & Iliescu, D. (2022). Secular changes in personality: An age-period-cohort analysis. *Journal of Research in Personality*, 100, 104280. <https://doi.org/10.1016/j.jrp.2022.104280>
- Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2021). Linking social-emotional learning to long-term success: Student survey responses show effects in high school and beyond. *Education Next*, 21(1), 64–71.
- Jackson, D. J. R., LoPilato, A. C., Hughes, D., Guenole, N., & Shalfrrooshan, A. (2016). The internal structure of situational judgement tests reflects candidate main effects:

- Not dimensions or situations. *Journal of Occupational and Organizational Psychology*, 90(1), 1–27.
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55(4), 243–252. <https://doi.org/10.1037/h0045996>
- Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, 44(4), 501–511. <https://doi.org/10.1016/j.jrp.2022.104280>
- Jackson, K. (2018). What do test scores miss? The importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5), 2072–2107.
- James, L. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, 1(2), 131–163. <https://doi.org/10.1177/109442819812001>
- John, O. P., & De Fruyt, F. (2015). *Education and social progress: Framework for the longitudinal study of social and emotional skills in cities* (EDU-CERI-CD(2015)13.en). OECD Publishing. [https://one.oecd.org/document/EDU/CERI/CD\(2015\)13/en/pdf](https://one.oecd.org/document/EDU/CERI/CD(2015)13/en/pdf)
- John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.
- Jones, S., Bailey, R., Brush, K., & Nelson, B. (2019). *Introduction to the taxonomy project: Tools for selecting & aligning frameworks*. CASEL. <https://measuringSEL.casel.org/wp-content/uploads/2019/02/Frameworks-C.1.pdf>
- Judge, T. A., & Higgins, C. (1998). Affective disposition and the letter of reference. *Organizational Behavior and Human Decision Processes*, 75, 207–221.
- Kane, M., Berryman, S., Goslin, D., & Meltzer, A. (1990). *The Secretary's Commission on Achieving Necessary Skills: Identifying and describing the skills required by work*. Pelavin Associates.
- Kankaraš, M. (2017). *Personality matters: Relevance and assessment of personality characteristics* (OECD Education Working Paper No. 157). OECD Publishing.
- Kankaraš, M., & Suarez-Alvarez, J. (2019). *Assessment framework of the OECD Study on Social and Emotional Skills* (OECD Education Working Papers 207). OECD Publishing. <https://doi.org/10.1787/5007adef-en>
- Kapteyn, A., Smith, J. P., van Soest, A. H. O., & Vonkova, H. (2011). *Anchoring vignettes and response consistency* (RAND Working Paper Series WR-840). RAND. <https://ssrn.com/abstract=1799563> or <http://dx.doi.org/10.2139/ssrn.1799563>
- Kasten, N., & Freund, P. A. (2016). A meta-analytical multilevel reliability generalization of Situational Judgment Tests (SJTs). *European Journal of Psychological Assessment*, 32(3), 230–240.
- Kasten, N., Freund, P. A., & Staufenbiel, T. (2018). “Sweet little lies”: An in-depth analysis of faking behavior on situational judgment tests compared to personality questionnaires. *European Journal of Psychological Assessment*, 36(1), 136–148. <http://dx.doi.org/10.1027/1015-5759/a000479>

- Kautz, T., & Zantoni, W. (2024). Measurement and development of noncognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal program. *Journal of Human Capital*, 18(2), 272–304. DOI: 10.1086/728087
- Keener, A. (2020). *A comparison of Cohen's Kappa and Gwet's AC1 with a mass shooting classification index: A study of rater uncertainty* [Doctoral dissertation]. Oklahoma State University, Stillwater.
- Kell, H. J., Martin-Raugh, M. P., Carney, L. M., Inglese, P. A., Chen, L., & Feng, G. (2017). *Exploring methods for developing behaviorally anchored rating scales for evaluating structured interview performance* (ETS Research Report RR-17-28). ETS.
- Khorrarnadel, L., Jeon, M., & Wang, L. L. (2019). Editorial: Advances in modeling response styles and related phenomena. *British Journal of Mathematical and Statistical Psychology*, 72(3), 393–400.
- Khorrarnadel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49(2), 161–177. <https://doi.org/10.1080/00273171.2013.866536>
- King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, 15(1), 46–66.
- Klieger, D. M., Kell, H. J., Rikoon, S., Burkander, K. N., Bochenek, J. L., & Shore, J. R. (2018). *Development of the behaviorally anchored rating scales for the skills demonstration and progression guide* (ETS Research Report RR-18-24). ETS.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805. <https://doi.org/10.1073/pnas.1218772110>
- Kubinger, K. D. (2009). The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behavior. *Acta Psychologica Sinica*, 41, 1024–1036.
- Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment*, 22, 101–107. <https://doi.org/10.1111/ijsa.12060>
- Kuncel N. R., & Tellegen A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, 62, 201–228.
- Kyllonen, P. C. (2006). *The research behind the ETS® Personal Potential Index. A background paper from ETS*. ETS. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.522.5309&rep=rep1&type=pdf>
- Kyllonen, P. C. (2013). Soft skills for the workplace. *Change. The Magazine for Higher Learning*, 45(6), 16–23.
- Kyllonen, P. C. (2015). *Developing definitions and assessment methods. National Academies of Sciences, Engineering, and Medicine* [Presentation]. Symposium on assessing hard-to-measure cognitive, interpersonal, and intrapersonal competencies, Washington, DC, United States.

- Kyllonen, P. C. (2016). Socio-emotional and self-management variables in learning and assessment. In A. Rupp & J. Leighton (Eds.), *Handbook of cognition and assessment* (pp.174–197). Wiley/Blackwell.
- Kyllonen, P. C., Heincke, P., Holtzman, S., Olivera-Aguilar, M., Iacovelli, M., Xu, J., Grodman, L., & DelMonico, B. (2020). *Predicting Yale School of Management outcomes with a forced-choice behavioral assessment* [Unpublished manuscript]. ETS.
- Kyllonen, P. C., & Bertling, J. P. (2014). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277–285). CRC Press.
- Kyllonen, P. C., & Kell, H. (2018). Ability tests measure personality, personality tests measure ability: Disentangling construct and method in evaluating the relationship between personality and ability. *Journal of Intelligence*, 6(3), 32. <https://doi.org/10.3390/jintelligence6030032>
- Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, 86(2), 602–640. <https://doi.org/10.3102/0034654315617832>
- LeBreton, J. M., Grimaldi, E. M., & Schoen, J. L. (2020). Conditional reasoning: A review and suggestions for future test development and validation. *Organizational Research Methods*, 23(1), 65–95. <https://doi.org/10.1177/1094428118816366>
- Legree, P. J., Kerner, B. S., & Shewach, O. R. (2021). *Identifying optimal keys to enhance personality scale validity* (Research Note 2021-04). U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lee, J., & Stankov, L. (2013). Higher-order structure of non-cognitive constructs and prediction of PISA 2003 math achievement. *Learning and Individual Differences*, 26, 119–130. <https://doi.org/10.1016/j.lindif.2013.05.004>
- Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39(2), 329–358. https://doi.org/10.1207/s15327906mbr3902_8
- Lee, P., Joo, S.-H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement*, 43(3), 226–240. <https://doi.org/10.1177/0146621618768294>
- Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, 2(1), 1–24.
- Levin, H. M. (1989). Ability testing for job selection: Are the economic claims justified? In B. R. Gifford (Ed.), *Test policy and the politics of opportunity allocation: The workplace and the law* (pp. 211–232). National Commission on Testing and Public Policy.
- Levin, H. M. (2012). More than just test scores. *Prospects: Quarterly Review of Comparative Education*, 42, 269–284.

- Likert, R. (1932). A technique for measurement of attitudes. *Archives of Psychology*, 140, 5–55.
- Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, 3(4), 486–512.
- Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and non-cognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, 3(1), 101–128.
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37(4), 426–441. <https://doi.org/10.1108/00483480810877598>
- Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology*, 94(4), 1095–1101.
- Lombardo, M. M., & Eichinger, R. W. (2004). *The leadership machine* (3rd ed.). Lominger International.
- Lubke, G. H., & Muthen, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, 11, 514–534.
- Ludlow, L., Anghel, E., Szendey, O., O’Keefe, T., Howell, B., Matz-Costa, C., & Braun, H. (2020). The Boston College Living a Life of Meaning and Purpose (BC-LAMP) portfolio: An application of Rasch/Guttman scenario methodology. *Journal of Applied Measurement*, 21(2), 134–153.
- Ludlow, L. H., Matz-Costa, C., & Klein, K. (2019). Enhancement and validation of the Productive Engagement Portfolio–Scenario (PEP–S8) Scales. *Measurement and Evaluation in Counseling and Development*, 52(1), 15–37. <https://doi.org/10.1080/07481756.2018.1497430>
- Madnani, N., & Cahill, A. (2018). *Automated scoring: Beyond natural language processing*. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1099–1109). Association for Computational Linguistics. <https://www.aclweb.org/anthology/C18-1094.pdf>
- Mahoney, J. L., Durlak, J. A., & Weissberg, R. P. (2018). An update on social and emotional learning outcome research. *Phi Delta Kappan*, 100(4), 18–23.
- Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., Ko, R., & Sanghvi, S. (2017). *Jobs lost, job gained: What the future of work will mean for jobs, skills, and wages*. McKinsey & Company. <https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages>
- Marsh, H. W., Hau, K.-T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD’s brief self-report measure of educational psychology’s most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6, 311–360.
- Martin-Raugh, M. P., Kyllonen, P. C., Hao, J., Bacall, A., Becker, D., Kurzum, C., Yang, Z., Yan, F., & Barnwell, P. (2020). Negotiation as an interpersonal skill: Generalizability

- of negotiation outcomes and tactics across contexts at the individual and collective levels. *Computers in Human Behavior*, 104, 105966. <https://doi.org/10.1016/j.chb.2019.03.030>
- Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 109–128). Routledge.
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, 45(6), 935–974. <https://doi.org/10.1080/00273171.2010.531231>
- McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. *Psychological Review*, 123(5), 569–591. <https://doi.org/10.1037/rev0000035>
- McAbee, S. T., & Oswald, F. L. (2013). The criterion-related validity of personality measures for predicting GPA: A meta-analytic validity competition. *Psychological Assessment*, 25(2), 532–544.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability: Monograph*. Rand Corporation.
- McCaffrey, D. F., Oliveri, M. E., & Holtzman, S. (2018). *A generalizability theory study to examine sources of score variance in third-party evaluations used in decision-making for graduate school admissions* (GRE Board Research Report No. GRE-18-03). ETS. <https://doi.org/10.1002/ets2.12225>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60, 63–91.
- McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86(4), 730–740.
- McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, 96(2), 327–336.
- McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence*, 33, 515–525. <https://doi.org/10.1016/j.intell.2005.02.001>
- Medvedev, O. N., Merry, A. F., Skilton, C., Gargiulo, D. A., Mitchell, S. J., & Weller, J. M. (2019). Examining reliability of WHOBARS: A tool to measure the quality of administration of WHO surgical safety checklist using generalizability theory with surgical teams from three New Zealand hospitals. *BMJ Open*, 9(1), e022625. <https://doi.org/10.1136/bmjopen-2018-022625>
- Meijer, R. R., & Niessen, A. S. M. (2015). A trial studying approach to predict college achievement. *Frontiers in Psychology*, 6, 887.
- Messick, S. (1978). *Potential uses of noncognitive measurement in education* (ETS Research Bulletin Series, 1, 1–25). ETS. <https://doi.org/10.1002/j.2333-8504.1978.tb01156.x>

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Meyer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, 59, 507–536. <https://doi.org/10.1146/annurev.psych.59.103006.093646>
- Meyer, R. H., Wang, C., & Rice, A. B. (2018). *Measuring students' social-emotional learning among California's CORE Districts: An IRT modeling approach* [Working paper]. Policy Analysis for California Education. https://www.edpolicyinca.org/sites/default/files/Measuring_SEL_May-2018.pdf
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Ministerio de Educacion Gobierno de Chile [Chilean Government Ministry of Education]. (2016). *Plan de evaluaciones nacionales e internacionales para el periodo 2016–2020* [National and international evaluation plan for the period 2016–2020]. Division Juridica, Solicitud No. 00197, Decreto No. 01820 (20 June 2016).
- Mirjafari, S., Masaba, K., Grover, T., Wang, W., Audia, P. G., Campbell, A. T., Chawla, N. V., Das Swain, V., De Choudhury, M., Dey, A. K., D'Mello, S. K., Gao, G., Gregg, J. M., Jagannath, K., Jiang, K., Lin, S., Liu, Q., Mark, G., Martinez, G. J., . . . Striegel, A. (2019). Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2), 1–24. ACM. <https://dl.acm.org/doi/10.1145/3328908>
- Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality & Quantity*, 42(6), 779–794. <https://doi.org/10.1007/s11135-006-9067-x>
- Moretti, E. (2005, March 1). Social returns to human capital. *The Reporter*, The National Bureau of Economic Research. <https://www.nber.org/reporter/spring-2005/social-returns-human-capital>
- Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, 40(7), 500–516. <https://doi.org/10.1177/0146621616662226>
- Möttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., Amoussou-Yéyé, D., Bäckström, M., Barkauskiene, R., Barry, O., Bhowon, U., Björklund, F., Bochaver, A., Bochaver, K., de Bruin, G., Cabrera, H. F., Chen, S. X., Church, A. T., Cissé, D. D., . . . Johnson, W. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin*, 38, 1423–1436. <https://doi.org/10.1177/0146167212451275>
- Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology*, 95(2), 321–333. <http://dx.doi.org/10.1037/a0017975>
- Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement*, 11(1), 1–31. <https://doi.org/10.1177/014662168701100101>

- Mumford, M. D., Whetzel, D. L., Murphy, S. T., & Eubanks, D. L. (2012). Background data. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 201–233). Routledge.
- Muraki, E., & Muraki, M. (2016). Generalized partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 127–138). Routledge.
- Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in academic admissions: A meta-analysis and cautionary tale. *College and University*, 84(4), 83–88.
- Muthén, L. K., & Muthén, B. O. (1998–2011). *Mplus User's Guide* (6th ed.). Muthén & Muthén.
- Naemi, B. D., Seybert, J., Robbins, S. B., & Kyllonen, P. C. (2014). *Examining the WorkFORCE™ Assessment for Job Fit and Core Capabilities of FACETS™* (ETS Research Report RR-14-32). <https://doi.org/10.1002/ets2.12040>
- National Academies of Sciences, Engineering, and Medicine. (2017). *Supporting students' college success: The role of assessment of intrapersonal and interpersonal competencies*. The National Academies Press. <https://doi.org/10.17226/24697>
- National Assessment Governing Board. (2002). *Background information framework for the National Assessment of Educational Progress*.
- National Assessment Governing Board. (2013). *Contextual information framework for the National Assessment of Educational Progress*. <https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/contextual-information/contextual-information-framework.pdf>
- National Association of Colleges and Employers. (2018). *2019 NACE job outlook report*. <https://www.odu.edu/content/dam/odu/offices/cmc/docs/nace/2019-nace-job-outlook-survey.pdf>
- National Center for Education Statistics. (n.d.-a). *National Education Longitudinal Study of 1988 (NELS:88): Questionnaires*. Institute of Education Sciences. <https://nces.ed.gov/surveys/nels88/questionnaires.asp>
- National Center for Education Statistics. (n.d.-b). *Survey questionnaires: What can survey questionnaires tell us about student achievement?* Institute of Education Sciences. https://nces.ed.gov/nationsreportcard/experience/survey_questionnaires.aspx
- National Research Council. (2008). *Research on future skill demands: A workshop summary*. The National Academies Press. <https://doi.org/10.17226/12066>.
- National Research Council (2011). *Assessing 21st century skills: Summary of a workshop*. The National Academies Press. <https://doi.org/10.17226/13215>
- National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. The National Academies Press. <https://doi.org/10.17226/13398>
- Nering, M., & Ostini, R. (Eds.) (2010). *Handbook of polytomous item response theory models*. Routledge.
- Neubauer, A. C., & Hofer, G. (2022). (Retest-)reliable and valid despite low alphas? An example from a typical performance situational judgment test of emotional

- management. *Personality and Individual Differences*, 189, 111511. <https://doi.org/10.1016/j.paid.2022.111511>
- Newstead, S. E., & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics*, 30(10), 1447–1462. <https://doi.org/10.1080/00140138708966038>
- Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment*, 13(4), 250–260. <https://doi.org/10.1111/j.1468-2389.2005.00322.x>
- Niessen, A. S. M., & Meijer, R. R. (2017). On the use of broadened admission criteria in higher education. *Perspectives on Psychological Science*, 12(3), 436–448. <https://doi.org/10.1177/1745691616683050>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting success in higher education using proximal predictors. *Plos ONE*, 11(4), e0153663.
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, 106, 183–189.
- Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93(1), 116–130.
- Novarese, M., & Di Giovinazzo, V. (2013). *Promptness and academic performance* (MPRA Paper No. 49746). University Library of Munich, Germany. <http://mpra.ub.uni-muenchen.de/49746/>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, 96(5), 966–980. <https://doi.org/10.1037/a0022955>
- O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L. III, & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment*, 15(1), 19–29. <https://doi.org/10.1111/j.1468-2389.2007.00364.x>
- O'Toole, B. I., & Stankov, L. (1992). Ultimate validity of psychological tests. *Personality and Individual Differences*, 13(6), 699–716.
- Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, 96(4), 762–773.
- Ohland, M. W., Loughry, M. L., Woehr, D. J., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Pomeranz, H. R., & Schmucker, D. G. (2012). The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self- and peer evaluation. *Academy of Management Learning & Education*, 11(4), 609–630. <http://dx.doi.org/10.5465/amle.2010.0177>

- Oostrom, J. K., De Soete, B., & Lievens, F. (2015). Situational judgment testing: A review and some new developments. In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment: Contemporary issues for theory and practice* (pp. 172–189). Routledge.
- Oostrom, J. K., Kobis, N. C., Ronay, R., & Cremers, M. (2017). False consensus in situational judgment tests: What would others do? *Journal of Research in Personality*, 71, 33–45.
- Organisation for Economic Co-operation and Development. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. <http://dx.doi.org/10.1787/9789264190511-en>
- Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Technical Report*. https://www.oecd.org/en/publications/pisa-2012-technical-report_6341a959-en.html
- Organisation for Economic Co-operation and Development. (2015). *Skills for social progress: The power of social and emotional skills*. OECD Publishing. <http://dx.doi.org/10.1787/9789264226159-en>
- Organisation for Economic Co-operation and Development. (2016). *PISA 2015 assessment and analytical framework: Science, reading, mathematics and financial literacy*. OECD Publishing. <https://doi.org/10.1787/9789264255425-1-en>
- Organisation for Economic Co-operation and Development. (2017a). *PISA 2015 database*. <https://www.oecd.org/pisa/data/2015database/>
- Organisation for Economic Co-operation and Development. (2017b). *PISA 2012 Technical Report*. https://www.oecd.org/en/publications/pisa-2012-technical-report_6341a959-en.html
- Organisation for Economic Co-operation and Development. (2019a). *OECD study on social and emotional skills*. <https://www.oecd.org/education/ceri/OECD-Study-on-Social-and-Emotional-Skills.pdf>
- Organisation for Economic Co-operation and Development. (2019b). *PISA 2018 assessment and analytical framework*. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Organisation for Economic Co-operation and Development. (2021). *Beyond academic learning: First results from the Survey of Social and Emotional Skills*. OECD Publishing <https://doi.org/10.1787/92a11084-en>
- Organisation for Economic Co-operation and Development. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://www.oecd.org/en/publications/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en.html
- Oswald, F. L., Behrend, T. S., & Foster, L. (Eds.). (2019). *Workforce readiness and the future of work*. Routledge.
- Oswald, F. L., & Schell, K. L. (2010). Developing and scaling personality measures: Thurstone was right—but so far, Likert was not wrong. *Industrial and Organizational Psychology*, 3(4), 481–484.
- Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of

- college student performance. *Journal of Applied Psychology*, 89(2), 187–207. <http://dx.doi.org/10.1037/0021-9010.89.2.187>
- Ortner, T., & Proyer, R. (2015). Objective personality tests (pp. 133–149). In T. M. Ortner & F. J. R. van de Vijver (Eds.), *Behavior-based assessment in psychology*. Hogrefe & Huber.
- Ortner, T. M., & Schmitt, M. (2014). Advances and continuing challenges in objective personality testing. *European Journal of Psychological Assessment*, 30(3), 163–168. <http://dx.doi.org/10.1027/1015-5759/a000213>
- Ozer, D., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401–421.
- Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice. *AMEE Guide No. 100*, 38(1), 3–17. <https://doi.org/10.3109/0142159X.2015.1072619>
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 67–88). Lawrence Erlbaum Associates.
- Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, 81(3), 524–539. <https://doi.org/10.1037/0022-3514.81.3.524>
- Pavlov, G., Shi, D., Maydeu-Olivares, A., & Fairchild, A. (2021). Item desirability matching in forced-choice test construction. *Personality and Individual Differences*, 183, 111114. <https://doi.org/10.1016/j.paid.2021.111114>
- Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, 65(1), 70–89.
- Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook of classification*. American Psychological Association; Oxford University Press.
- Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1999). *An occupational information system for the 21st century: The development of O*NET*. American Psychological Association. <https://doi.org/10.1037/10313-000>
- Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment*, 11(1), 1–16. <https://doi.org/10.1111/1468-2389.00222>
- Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, 7, 27–65. <http://dx.doi.org/10.1177/1094428103259554>
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135(2), 322–338. <https://doi.org/10.1037/a0014996>
- Poropat, A. (2014). Other-rated personality and academic performance: Evidence and implication. *Learning and Individual Differences*, 34, 24–32.

- Prasad, J. (2017). *It's both who you are and where you're from: Relating vocational interests and socioeconomic status to bias in biodata and SJTs* [Master's thesis, Michigan State University]. <https://d.lib.msu.edu/etd/6934>
- Prasad, J. J., Showler, M. B., Ryan, A. M., Schmitt, N., & Nye, C. D. (2017). When belief precedes being: How attitudes and motivation before matriculation lead to fit and academic performance. *Journal of Vocational Behavior*, 100, 27–42. <https://doi.org/10.1016/j.jvb.2017.02.003>
- Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement*, 71(3), 523–550
- Primi, R., Zanon, C., Santos, D., De Fruyt, F., & John, O. P. (2016). Anchoring vignettes: Can they make adolescent self-reports of social-emotional skills more reliable, discriminant, and criterion-valid? *European Journal of Psychological Assessment*, 32, 39–51. <https://doi.org/10.1027/1015-5759/a000336>
- Primi, R., Santos, D., De Fruyt, F., & John, O. P. (2019). Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal of Mathematical and Statistical Psychology*, 72(3), 447–465. <https://doi.org/10.1111/bmsp.12168>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611–630.
- Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, 55(4), 361–382.
- Rath, T. (2007). *StrengthsFinder 2.0*. Gallup Press.
- Rath, T., & Conchie, B. (2008). *Strengths based leadership: Great leaders, teams, and why people follow*. Gallup Press.
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.R-project.org/>
- Reeves, R. V., & Venator, J. (2014). *Jingle-jangle fallacies for non-cognitive factors*. Brookings. <https://www.brookings.edu/articles/jingle-jangle-fallacies-for-non-cognitive-factors/>
- Reise, S. P., & Revicki, D. A. (Eds.). (2015). *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.
- Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, 43(1), 73–97. <https://doi.org/10.1177/0049124113509605>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138, 353–387.
- Rieger, S., Göllner, R., Spengler, M., Trautwein, U., Nagengast, B., & Roberts, B. W. (2017). Social cognitive constructs are just as stable as the Big Five between grades 5 and 8. *AERA Open*, 3(3). <https://doi.org/10.1177/2332858417717691>

- Rimfeld, K., Malanchini, M., Krapohl, E., Hannigan, L. J., Dale, P. S., & Plomin, R. (2018). The stability of educational achievement across school years is largely explained by genetic factors. *Science of Learning*, 16(3), 1–10. <https://doi.org/0.1038/s41539-018-0030-0>
- Rios, J. A., Ling, G., Pugh, R., Becker, D. M., & Bacall, A. N. (2020). Identifying critical 21st century for workplace success: A content analysis of job advertisements. *Educational Researcher*, 49, 80–89.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261–288.
- Robbins, S. B., Oh, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and persistence, mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *Journal of Applied Psychology*, 94, 1163–1184.
- Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, 43, 137–145. <https://doi.org/10.1037/0033-2909.132.1.1>
- Roberts, B. W., Kuncel, N., Shiner, R. N., Caspi, A., & Goldberg, L. (2007). The power of personality: A comparative analysis of the predictive validity of personality traits, SES, and IQ. *Perspectives in Psychological Science*, 2, 313–345.
- Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, 143(2), 117–141. <https://doi.org/0.1037/bul0000088>
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, 24(1), 3–32. <https://doi.org/10.1177/01466216000241001>
- Rothmann, S., & Coetzer, E. P. (2003). The Big Five personality dimensions and job performance. *Journal of Industrial Psychology*, 29, 68–74.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 80(1), 1–28. <https://doi.org/10.1037/h0092976>
- Salgado, J. F. (1997). The five-factor model of personality and job performance in the European community. *Journal of Applied Psychology*, 82, 30–43.
- Salgado, J. F., & Lado, M. (2018). Faking resistance of a quasi-ipsative forced-choice personality inventory without algebraic dependence. *Journal of Work and Organizational Psychology*, 34(3), 213–216. <https://doi.org/10.5093/jwop2018a23>
- Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, 23(1), 3–30. <https://doi.org/10.1080/1359432X.2012.716198>
- Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups:

- A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, 88(4), 797–834.
- Salzburg Global Seminar. (2018). *The Salzburg statement for social and emotional learning*. https://www.salzburgglobal.org/fileadmin/user_upload/Documents/2010-2019/2018/Session_603/SalzburgGlobal_Statement_SEL.pdf
- Samejima, F. (2016). Graded response model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 77–108). Routledge.
- Sass, R., Frick, S., Reip, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivational forced-choice versus rating scale instruments. *Assessment*, 27(3), 572–584. <https://doi.org/10.1177/1073191118762049>
- Saucier, G. (2000). Isms and the structure of social attitudes. *Journal of Personality and Social Psychology*, 78(2), 366–385. <https://doi.org/10.1037/0022-3514.78.2.366>
- Saucier, G. (2013). Isms dimensions: Toward a more comprehensive and integrative model of belief-system components. *Journal of Personality and Social Psychology*, 104, 921–939.
- Schaeppers, P. C., Freudenstein, J.-P., Mussel, P., Lievens, F., & Krumm, K. (2020). Effects of situation descriptions on the construct-related validity of construct-driven situational judgment tests. *Journal of Research in Personality*, 87, 1–5.
- Schanzenbach, D. W., Nunn, R., Bauer, L., Mumford, M., & Breitwieser, A. (2016). *Seven facts on noncognitive skills from education to the labor market*. The Hamilton Project: https://www.hamiltonproject.org/assets/files/seven_facts_noncognitive_skills_education_labor_market.pdf
- Schläpfer, F., & Fischhoff, B. (2010). *When are preferences consistent? The effects of task familiarity and contextual cues on revealed and stated preferences* (Working Paper No. 1007). University of Zurich, Socioeconomic Institute. <https://www.econstor.eu/bitstream/10419/76190/1/638617693.pdf>
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schmitt, N., & Ali, A. A. (2015). The practical importance of measurement invariance. In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 327–346). Routledge.
- Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, 94(6), 1479–1497.
- Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review*, 1, 173–208.
- Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8), 1438–1457.
- Segal, C. (2013). Misbehavior, education, and labor market outcomes. *Journal of the European Economic Association*, 11(4), 743–779. <https://doi.org/10.1111/jeea.12025>

- Seybert, J., & Becker, D. (2019). *Examination of the test-retest reliability of a forced-choice personality measure* (ETS Research Report RR-19-37). ETS.
- SHL Group. (1999). *OPQ32: Manual and user's guide*.
- Shultz, M. M., & Zedeck, S. (2011). Predicting lawyering effectiveness: Broadening the basis for law school admission decisions. *Law & Social Enquiry*, 36(3), 620–661. <https://doi.org/10.1111/j.1747-4469.2011.01245.x>
- Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4), 549–571. <https://doi.org/10.1177/0956797617739704>
- Slavich, G. M., Taylor, S., & Picard, R. W. (2019). *Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations*. *Stress*, 22(4), 408–413. <https://doi.org/10.1080/10253890.2019.1584180>
- Society for Human Resource Management. (2014). *SHRM Body of Competency and Knowledge™*. <https://snv.shrm.org/sites/snv.shrm.org/files/SHRMBoCK.pdf>
- Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st century competencies: Guidance for educators*. Asia Society. <https://asiasociety.org/files/gcen-measuring21cskills.pdf>
- Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: *The BESSI Journal of Personality and Social Psychology*, 123(1), 192–222. <https://doi.org/10.1037/pspp0000401>
- Stankov, L., Morony, S., & Lee, Y.-P. (2013). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology*, 34(1), 9–28. <https://doi.org/10.1080/01443410.2013.814194>
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29(3), 184–203.
- Stecher, B., & Hamilton, L. (2014). *Measuring hard-to-measure student competencies: A research and development plan*. The RAND Corporation.
- Stemler, S. E., Bebell, D., & Sonnabend, L. A. (2011). Using school mission statements for reflection and research. *Educational Administration Quarterly*, 47(2), 383–420. <https://doi.org/10.1177/0013161X10387590>
- Sternberg, R. J., & Horvath, J. A. (1999). *Tacit knowledge in professional practice: Researcher and practitioner perspectives*. Lawrence Erlbaum Associates.
- Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. Cambridge University Press.
- Stone, T. H., Foster, J., Webster, B. D., Harrison, J., & Jawahar, I. M. (2016). Gender differences in supervisors' multidimensional performance ratings: Large sample evidence. *Human Performance*, 29(5), 428–446. <http://dx.doi.org/10.1080/08959285.2016.1224884>

- Stricker, L. J., & Rock, D. A. (1998). Assessing leadership potential with a biographical measure of personality traits. *International Journal of Selection and Assessment*, 6(3), 164–184. <https://doi.org/10.1111/1468-2389.00087>
- Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling*, 27(1), 111–130. <https://doi.org/10.1080/10705511.2019.1602776>
- Taylor, R. D., Oberle, E., Durlak, J. A., Weissberg, R. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, 88(4), 1156–1171.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44, 703–742.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567–577
- Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25–29.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–286. <https://doi.org/10.1037/h0070288>
- Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Wiley.
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9(1), 151–176.
- U.S. Office of Personnel Management. (2018a). *Situational judgment tests*. <https://www.opm.gov/policy-data-oversight/assessment-and-selection/other-assessment-methods/situational-judgment-tests/>
- U.S. Office of Personnel Management. (2018b). *Assessment & selection*. <https://www.opm.gov/policy-data-oversight/assessment-and-selection/>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory*. Chapman and Hall/CRC.
- Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality & Social Psychology*, 98(2), 281–300.
- Vedel, A. (2014). The Big Five and tertiary academic performance: A systematic review and meta-analysis. *Personality and Individual Differences*, 71, 66–76. <https://doi.org/10.1016/j.paid.2014.07.011>
- von Davier, M. (2016). Rasch model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 1, Models* (pp. 31–50). Routledge.
- von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*, 42(4), 291–306. <https://doi.org/10.1177/0146621617730389>

- Vukasovic, T., & Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin*, 141(4), 769–785. <https://doi.org/10.1037/bul0000017>
- Wand, J., King, G., & Lau, O. (2011). Anchors: Software for anchoring vignettes data. *Journal of Statistical Software*, 42(3), 1–25.
- Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, 41(8), 600–613. <https://doi.org/10.1177/0146621617703183>
- Wanzer, D., Postlewaite, E., & Zargarpour, N. (2019). Relationships among noncognitive factors and academic performance: Testing the University of Chicago Consortium on School Research Model. *AERA Open*, 5(4), 1–20. <https://doi.org/10.1177/2332858419897275>
- Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, 29(7), 1159–1177. <https://doi.org/10.1177/0956797618761661>
- Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics*, 96(4), 849–861.
- Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Bohme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, 41(2), 115–129. <https://doi.org/10.1177/0146621616676791>
- Weng, Q., Yang, H., Lievens, F., & McDaniel, M. A. (2018). Optimizing the validity of situational judgment tests: The importance of scoring methods. *Journal of Vocational Behavior*, 104, 199–209.
- West, M. R. (2016). *Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts*. Brookings. <https://www.brookings.edu/research/should-non-cognitive-skills-be-included-in-school-accountability-systems-preliminary-evidence-from-californias-core-districts/>
- West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology*, 55, 119–129. <https://doi.org/10.1016/j.appdev.2017.06.001>
- West, M. R., Pier, L., Fricke, H., Loeb, S., Meyer, R. H., & Rice, A. B. (2018). *Trends in student social-emotional learning: Evidence from the CORE Districts* (Working paper). PACE: Policy Analysis for California Education, CORE-PACE Research Partnership. <https://edpolicyinca.org/publications/trends-student-social-emotional-learning>
- Westring, A. J. F., Oswald, F. L., Schmitt, N., Drzakowski, S., Imus, A., Kim, B., & Shivpuri, S. (2009). Estimating trait and situational variance in a situational judgment test. *Human Performance*, 22(1), 44–63. <https://doi.org/10.1080/08959280802540999>
- Whelpley, C. E. (2014). *How to score situational judgment tests: A theoretical approach and empirical test* [Doctoral dissertation, Virginia Commonwealth University]. <https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=4607&context=etd>

- Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19(3), 188–202. <https://doi.org/10.1016/j.hrmr.2009.03.007>
- Wise, S. L., & Gao, L. (2014). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343–354. <https://doi.org/10.1080/08957347.2017.1353992>
- Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, 19, 373–390.
- Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, 81(4), 1014–1045. <https://doi.org/10.1007/s11336-016-9506-0>
- Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573, 364–369. <https://doi.org/10.1038/s41586-019-1466-y>
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040. <https://doi.org/10.1073/pnas.1418680112>
- Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics*, 72(C), 51–60. <https://doi.org/10.1016/j.socec.2017.11.005>
- Zu, J., & Kyllonen, P. C. (2020). Nominal response model is useful for scoring multiple-choice situational judgment tests. *Organizational Research Methods*, 23(2), 342–366. <https://doi.org/10.1177/1094428118812669>

NOTES

1. Selection due to death could explain some of these findings, as it has with cognitive ability (O'Toole & Stankov, 1992).
2. Meta-analyses typically report both a raw Spearman correlation (r) and a disattenuated correlation adjusted for measurement error in the predictor (ρ).
3. Alternatively, standardized alpha is based on average interitem correlation rather than covariance. C. F. Falk and Savalei (2011) discussed the differences in meaning between the two alphas with respect to personality measures. If a composite is the sum of the raw item scores because items are on the same scale, alpha is justified; if a composite is the sum of standardized item scores, as a result of items being on different scales, standardized alpha is justified.
4. Other more general IRT treatments can be found in Embretson and Reise (2000) and de Ayala (2009). Specialized treatments covering polytomous (multiple-category) IRT models can be found in Thissen and Wainer (2001) and Nering and Ostini (2010).

Large-Scale Assessments in International Contexts

Henry I. Braun
Boston College

Irwin Kirsch
ETS (Retired)

Broadly defined, international large-scale assessments (ILSAs) are survey-based studies designed to assess the knowledge, skills, experiences, behaviors, and attitudes of populations in an international, comparative context. Depending on the population of interest, ILSAs can be characterized as school based or household based. School-based ILSAs select participants (e.g., students, teachers, school principals) based on their school membership and location, whereas household-based ILSAs select participants based on the location of their households.

For school-based populations, ILSAs produce estimates of distributions of learning outcomes in key domains for subpopulations defined by various combinations of individual characteristics and contextual factors. The data collected by school-based ILSA enable estimation of the strength of the relationships between learning outcomes and students' backgrounds and school-related factors. Moreover, administering an ILSA periodically enables tracking over time of learning outcomes, as well as changes in the patterns of relationships of learning outcomes to selected individual characteristics and background factors. Analogous statements pertain to household-based studies, usually referred to as surveys of adult populations, with the modification that the focus on school characteristics is replaced with educational, social, and labor market outcomes. In addition, because adult surveys encompass a broad age range, they enable comparisons across age cohorts, yielding useful information on secular trends.

Today's ILSAs can trace their origins to pilot studies that began in the late 1950s and, by the early 1960s, had led to the creation of the International Association for the Evaluation of Educational Achievement (IEA). From a methodological perspective, ILSAs are also beneficiaries of efforts in the United States during the 1960s that culminated in the National Assessment of Educational Progress (NAEP). Prior to these initiatives, no systematic or standardized comparative effort had been undertaken to determine what different student populations knew or could do in various cognitive domains, including language arts and mathematics. This newfound attention to educational outcomes marked a major enhancement of reporting efforts that heretofore had focused primarily on educational inputs, such as the number of schools, teachers, and students in a particular country and how these numbers changed over time.¹

The sponsors and developers of IEA assessments and of NAEP believed that the ongoing, systematic collection and analysis of outcome data, as well as patterns of relationships between those outcomes and various factors, would yield valuable insights. The interest in using the information generated by large-scale assessments to track trends in outcomes and to acquire deeper understandings about the relationships between outcomes on the one hand and individual or contextual characteristics on the other hand marked a milestone in an evidence-based approach to educational policy. Over time, the complex dynamics between ILSAs and the ever-changing education policy landscape have had important implications for the methodologies used to develop and deploy these assessments, as well as for data analysis and reporting results, topics we will return to later in this chapter. It bears mentioning that a key feature of ILSAs is that they always report score distributions and patterns of relationships at a group level,

in contrast to traditional testing programs that report individual-level results. This feature also has important implications for the design and analysis of ILSAs and is the reason why such large-scale assessments are often referred to as group-score assessments.

Since their inception, ILSAs have experienced remarkable growth in participation and salience, reflecting, we believe, perceptions of increased utility of comparative information among policy makers and key stakeholders. Perceived utility is fueled in large part by mounting concerns about the levels and distributions of human capital² and how they are associated with longer term outcomes for individuals and societies. Of course, the actual utilization of the information gleaned from ILSAs depends on a host of political, economic, and other factors that vary across countries, as well as within countries, over time (Feuer, 2012, 2013; Ritzen, 2013).

At the same time, because of the cross-sectional nature of ILSAs, there are limitations to the inferences one can make regarding the relative efficacy of different educational systems in building human capital as well as the consequences for individuals as they make their way through the educational system and then later transition into adulthood and participate in the labor market (Singer & Braun, 2018). Although ILSA outcomes, especially league tables for school-level results, generate headlines and lead to calls for action, the road from evidence to constructive policy impact is strewn with potholes. In particular, because of their cross-sectional approach to data collection, attempts to employ ILSA data to establish causal linkages are at best speculative and at worst misleading. Unfortunately, these limitations are not always understood or respected in policy discussions. The increased visibility of ILSA results along with occasional, ILSA-based, overly broad prescriptions for education reforms have given rise to criticisms of the influence of ILSAs on education policy—particularly with regard to the “homogenization” of national education systems (e.g., Benavot & Smith, 2020; Meyer & Benavot, 2013). A related criticism concerns the applicability of education policies in one nation to the systems of another—especially if the nations differ markedly with respect to history, culture, and governance (Carnoy et al., 2015). These concerns are addressed in this chapter’s section on “Policy and Political Challenges.”

In the case of adult surveys, interest in the results is distributed across many governmental departments, each being concerned with one or another aspect of the study. We can expect that as a consequence of pandemic-related disruptions, there will be increased attention to the need for upskilling or reskilling. Concerns about the use of data for policy in this realm appear to be milder than those in the realm of K–12 education policy.

In the early 21st century, ILSA results enter policy landscapes characterized by complex interactions among multiple stakeholders with access to a variety of sources of information. Accordingly, it is not a simple matter to evaluate whether changes in ILSA instrumentation and procedures are enhancing the value of the assessment enterprise as a source of policy-relevant information. To this task, we advance the notion that a deeper understanding of the utility of ILSAs can be obtained through consideration of a framework introduced by Messick (1987). Messick argued that large-scale assessments

were a form of policy research and, accordingly, should be judged by their contributions to policy analysis and policymaking. His framework was intended to decompose policy utility into its constituent components to provide clearer guidance to assessment sponsors and assessment designers. In this chapter, Messick's framework is refined and extended to take account of more than 30 years of further development and, especially, the transition to digitally-based technologies. Employing this framework, we argue that digital technologies have the potential to further increase the relevance and utility of ILSAs, contributing to their continued growth.

Increased participation in ILSAs, along with the greater prominence of their results, has led to a number of challenges. The task of accommodating increased heterogeneity in the distributions of proficiency and in the languages of the assessments, within rigid constraints of time and budget, has heretofore been accomplished through significant process innovation and methodological advances (as will be detailed in the section "Transitioning to Digitally-Based Assessments"). For example, countries can choose in which language or languages they want to assess and report results. For school-based surveys, these results are typically related to the language(s) of instruction, whereas for adult assessments, they are related to the language(s) of the society. Nonetheless, there are surely limits that will soon be reached and alternative strategies will be required.

A second challenge is that increasing heterogeneity among participating countries has given rise to concerns that cross-national comparability has been compromised to some degree. More specifically, in the case of school-based ILSAs, comparative analyses are complicated by the fact that, across countries, substantially different proportions of a birth cohort are enrolled in school and, of those enrolled, not all have been exposed to the same educational experiences, varying both in quantity and in quality. In the case of adult surveys, different proportions of the target populations are accessible through household surveys. Helping stakeholders understand the implications for policy analysis of these challenges to comparability remains a work in progress.

Following this introduction, we provide a brief description of several key large-scale student and adult assessments. Then we place these assessments in their historical context noting two key inflection points—the first marked by significant innovations in measurement and psychometrics and the second by significant innovations in administration and data collection resulting from the introduction of digital platforms. With the emergence of these assessments as significant sources of credible evidence for policy makers and other key stakeholders, we next introduce and extend a framework proposed by Messick (1987) that provides both developers and users of ILSAs with a set of design criteria to evaluate the potential utility of these surveys.

With the transition to digital platforms, we then focus on the key role of digital platforms in the management, development, and implementation of ILSAs and how new technologies are impacting their workflow and related processes. Next we address methodological advances impacting the scaling and analysis of the data, and also the production and dissemination of different data products. Then we focus on a range of issues concerning ways to evaluate the impact of ILSAs and some key technical

challenges that must be addressed in future cycles, as well as a number of the political challenges their growth and influence have raised. We conclude with a summary and final conclusions.

OVERVIEW OF KEY INTERNATIONAL ASSESSMENTS

A major focus of this chapter is on the three major school-based ILSAs: Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Programme for International Student Assessment (PISA). The first two are sponsored by IEA; the last is sponsored by the Organisation for Economic Co-operation and Development (OECD). However, it also describes a number of other ILSAs covering a range of domains and populations.

These assessments each include two sets of instruments: one for measuring cognitive outcomes and one for eliciting information on a range of background characteristics and contextual factors. Cognitive outcomes include the core domains of reading, literacy, mathematics, and science knowledge and skills, as well as other domains that are assessed with less frequency or are introduced as innovative domains (e.g., problem-solving, financial literacy). Instruments for eliciting background characteristics are targeted to the participating students, their teachers, school principals, and their parents. It bears mentioning that the IEA assessments employ a curriculum-based model for defining their assessment frameworks. In contrast, PISA employs more of a literacy-based model. Although these studies may appear to be very similar, deep conceptual differences manifest in many ways, including sample selection, item development, and instrument design. Table 20.1 contains information on each of these assessments, providing a brief description of content area, numbers of participants, and years conducted.

TIMSS is administered on a 4-year cycle to students in Grades 4 and 8.³ Students are assessed in both mathematics and science. Participating countries include countries, subnational jurisdictions (e.g., states and provinces), and even school districts. In 2019, TIMSS began a transition to computer-based assessment delivery. About half the countries used digital devices for the assessment. TIMSS 2023 completed the transition to computer-based assessment, facilitating the introduction of innovative item types for more comprehensive coverage of the TIMSS mathematics and science frameworks. Scores in each TIMSS cycle are placed on a scale originally established in 1995.

PIRLS is a literacy assessment administered to fourth graders on a 5-year cycle.^{4,5} In 2016, the IEA introduced ePIRLS, a computer-based assessment of online reading. PIRLS participants had the option of also participating in a pilot test of ePIRLS. For participating countries, students typically took PIRLS one day and ePIRLS on the following day.⁶ For the 2016 administration, the IEA also introduced PIRLS Literacy, a reading assessment intended for countries with lower levels of reading proficiency. Through a linking strategy based on common items, scores on this assessment are placed on the main PIRLS scale.

Table 20.1 School-Based International Tests of Educational Achievement: Scope and Timing

Sponsor ^a	Description	Countries	Year(s) Conducted
IEA	First International Mathematics Study (FIMS)	12 countries	1964
IEA	Six Subject Study		1970–1971
	Science	19 systems	
	Reading	15 countries	
	Literature	10 countries	
	French as a foreign language	8 countries	
	English as a foreign language	10 countries	
	Civic education	10 countries	
IEA	First International Science Study (FISS; part of Six Subject Study)	19 countries	1970–1971
IEA	Second International Mathematics Study (SIMS)	10 countries	1982
IEA	Second International Science Study (SISS)	19 systems	1983–1984
ETS	First International Assessment of Educational Progress (IAEP-I, Mathematics Study and Science)	6 countries (12 systems)	1988
ETS	Second International Assessment of Educational Progress (IAEP-II, Mathematics and Science)	20 countries	1991
IEA	Reading Literacy (RL)	32 countries	1990–1991
IEA	Computers in Education	22 countries	1988–1989
		12 countries	1991–1992
IEA	International Computers and Information Literacy Study	35 countries	2013
		13 countries	2018
		30 countries	2023
IEA	Preprimary Project:		
	Phase I	11 countries	1989–1991
	Phase II	15 countries	1991–1993
	Phase III (longitudinal follow-up of Phase II sample)	15 countries	1994–1996
IEA	Third International Mathematics and Science Study (TIMSS)	45 countries	1994–1995
		40 countries	1997–1998

Sponsor ^a	Description	Countries	Year(s) Conducted
IEA	Civics Education Study (ICCS)	28 countries	1999
		35 countries	2009
		24 countries	2016
		25 countries	2022
OECD	Program for International Student Assessment (PISA) ^b	43 countries	2000 (reading)
		41 countries	2003 (math)
		57 countries	2006 (science)
		65 countries	2009 (reading)
		64 countries	2012 (math)
		72 countries	2015 (science)
		79 countries	2018 (reading)
		About 88 countries	2022 (math)
IEA	Progress in International Reading Literacy Study (PIRLS)	34 countries	2001
		41 countries	2006
		48 countries	2011
		50 countries	2016
		About 70 countries	2021
IEA	Trends in International Mathematics and Science Study (TIMSS)	45 countries	2003
		48 countries	2007
		63 countries	2011
		57 countries	2015
		64 countries	2019
		About 65 countries	2023

Note. Adapted from “Sampling Issues in Design, Conduct, and Interpretation of International Comparative Studies of School Achievement” by J. R. Chromy, in A. C. Porter and A. Gamoran (Eds.), *Methodological Advances in Cross-National Surveys of Educational Achievement* (pp. 80–117, Table 4.1), 2002, National Academies Press.

^a IEA = International Association for the Evaluation of Educational Achievement; OECD = Organisation for Economic Co-operation and Development.

^b Subject in parentheses is the focal domain for that administration.

PISA assesses 15-year-old students in the core domains of reading, mathematics, and science literacy on a 3-year cycle. In each country, eligible students are those enrolled in Grade 7 or above. During each cycle, one of the three core domains is the focal domain and therefore more testing time (i.e., items) is devoted to that domain than to the others. During the 2015 cycle, PISA completed the transition from a paper-based assessment to a computer-based assessment while maintaining a limited

paper-based option for countries not selecting the computer-based option. During the 2018 cycle, PISA also introduced multistage, adaptive testing as part of the computer-based assessment. In PISA 2018, each student was tested in two of the three core domains. As options for countries participating in the computer-based assessment, PISA currently assesses an innovative domain beyond the three core domains as well as the domain of financial literacy. Beginning with the 2025 cycle of PISA there will be an optional foreign language assessment that will alternate cycles with the optional financial literacy assessment.

In addition to these school-based surveys, this chapter also discusses one major assessment of adult skills, the Programme for the International Assessment of Adult Competencies (PIAAC), sponsored by the OECD. The assessment of adult literacy began with a number of national assessments in the United States.⁷ The design of the International Adult Literacy Survey (IALS), conducted in three rounds between 1994 and 1999, drew on the frameworks developed and refined by these earlier national surveys. Information on IALS and subsequent international surveys of adult literacy can be found in Figure 20.1.

PIAAC, first fielded in 2012, is a computer-based assessment for adults ages 16–65 in literacy and numeracy and a measure of problem-solving in technology-rich environments. This household survey was administered in three rounds over a period of 5 years (2012, 2015, and 2017). These rounds enabled more countries to participate within

	Population	Assessment Domains	Mode
IALS 1994–1998	Ages 16–65 19 countries Three rounds (1994, 1996, 1998)	Prose, Document & Quantitative Literacy	Paper & Pencil Open-Ended Response
All 2003–2008	Ages 16–65 11 countries Two rounds (2003, 2008)	Prose & Document Literacy, Numeracy, Problem-Solving/Analytical Reasoning	Paper & Pencil Open-Ended Response
PIAAC 2012–2024	Ages 16–65 Cycle 1: 2012 38 countries Three rounds (2012, 2014, 2017) Cycle 2: 2024 31 countries	Cycle 1: Literacy, Numeracy, Reading Components, & Problem-Solving in Technology Rich Environments Cycle 2: Literacy, Numeracy, Adaptive Problem-Solving, Reading & Numeracy Components	Computer-Based Assessment Adaptive Simulation Tasks Cycle 1: Laptop delivery Cycle 2: Tablet delivery

FIGURE 20.1

International Adult Assessments

a cycle and allowed countries (e.g., the United States) to do further national studies between cycles.

Laptops were used for data collection in all three rounds. For those respondents who were unwilling or unable to take the assessment on a laptop computer, a paper-and-pencil option was available for the assessment of literacy and numeracy only. In addition to the cognitive instruments, PIAAC incorporates an extensive background questionnaire, capturing information on a broad range of issues including demographics, education, labor force status, work experience, occupation, use of cognitive skills at home and at work, and incomes.

PIAAC is developed and implemented on a 10-year cycle. At the time of writing, PIAAC is conducting Cycle 2 with 31 participating countries, all of which are employing tablets for survey delivery.

It should be noted that in addition to the well-known global assessments that are discussed in this chapter, a number of other international (but regional) assessments play important roles in regional conversations related to education by providing information for credible cross-national comparisons. Information about those assessments can be found in a recent World Bank report (Clarke & Luna-Bazaldua, 2021).

Finally, the IEA also sponsors two periodic, specialized assessments: the International Computer and Information Literacy Study (ICILS) and International Civic and Citizenship Education Study (ICCS). The ICILS instrument assesses students in the eighth grade (or its equivalent) with regard to their computer literacy and information literacy. An optional module assesses computational thinking. Through the collection of additional background information, ICILS reports describe how these components of digital competence relate not only to one other, but also to relevant school and out-of-school contexts.⁸

The ICCS assesses students in the eighth or ninth grades. It reports on students' knowledge and understanding of concepts and issues related to civics and citizenship, as well as their beliefs, attitudes, and behaviors with respect to this domain. In addition, ICCS collects rich contextual data on the organization and content of civic and citizenship education in the curriculum, teacher qualifications and experiences, teaching practices, school environment and climate, and home and community support.⁹

HISTORY OF INTERNATIONAL ASSESSMENTS

Comparative international assessments had their genesis in a pilot study involving a mathematics assessment in six countries in the late 1950s, followed by a 12-country study, comprising five subjects, conducted by the IEA in 1964 (Husén, 1967; Suter, 2019). These early studies were motivated by the recognition that the education systems and policies of different countries constituted a natural laboratory, and by using a common assessment, international comparisons could yield insights on what strategies and

policies deserved further study. By the same token, the structure of the accompanying background questionnaire was motivated more by models of learning than by considerations of education policy.

Subsequently, the IEA sponsored a series of studies through the 1970s and 1980s, focusing primarily on mathematics and science (assessed separately), culminating in the Third International Mathematics and Science Study (TIMSS)¹⁰ in 1995 with 45 participating countries. Building on its Six Subjects Study (1970–1971) and its Reading Literacy Study (1990–1991), the IEA introduced PIRLS in 2001.

It is remarkable that in the late 1980s, in anticipation of the growing importance of international assessments, a number of U.S. federal agencies established, under the auspices of the National Academy of Sciences, the Board on International and Comparative Studies in Education (BICSE).¹¹ The main focus of BICSE was to evaluate the technical quality of these studies and to assess their policy relevance—all, presumably, to provide high-level guidance on how to enhance both (Bradburn & Gilford, 1990). As described by Heyneman and Lee (2014, p. 41), BICSE made a strong case for the utility of international studies. Echoing the original impetus behind such studies, it argued that variations in policies and practices among countries provided a natural laboratory in which to study the consequences of such differences. Further, BICSE noted the value to the United States of cross-national comparisons that could bring to light new concepts or an empirical basis for challenging long-held assumptions. These considerations have retained their cogency over the ensuing decades.¹²

The OECD introduced PISA in 2000 and PIAAC in 2012.¹³ PIAAC built on two earlier assessments of adults ages 16–65: IALS, which was administered in multiple rounds from 1994 to 1999, and the survey of Adult Literacy and Lifeskills, administered between 2003 and 2008 (see Kirsch et al., 2017, for a history of adult assessments). As will be elaborated, PIAAC is distinguished by the fact that it was the first ILSA to be designed from the outset to be delivered by computer. To fully appreciate the evolution of the systems that support ILSA efforts, however, it is informative to review the history of NAEP.

The history of NAEP has been addressed at length elsewhere (Beaton & Barone, 2017; Jones & Olkin, 2004). Building on substantial planning and development work, the first NAEP assessment was conducted in 1969. Bowing to political constraints related to the primacy of states and local districts in matters of education, the NAEP team had to ensure that the assessment design and sampling plan were such that no summative test scores could be developed, no individual-level results could be reported, and no population groups could be described at the state or local level. This mandate was further reinforced by the decision to report results only at the item level, with each item connected to a particular learning objective so that it could be judged for relevance by both professional educators and individual citizens.

Over the next decade, the context for education and education policy evolved considerably—resulting in a greater federal role in education alongside a growing concern that many schools across the country were neither serving well the needs

of particular subgroups nor accommodating changes in society and labor markets. Policy makers and other key stakeholders began to argue for the need for the kinds of information that the original NAEP design was unable to provide. This led to the establishment of a national commission with the goal of evaluating NAEP and identifying its strengths and weaknesses. Its report (Wirtz & Lapointe, 1982) emphasized the limitations of the existing NAEP design with respect to the interpretability and utility of the results. Drawing on this report, researchers from ETS developed an approach that would enable a new NAEP to address a set of important questions that included the following:

- Are all students learning the skills and flexibility they will need for their own and society's well-being in the 1980s and beyond?
- Are all students being well prepared, and do they have similar opportunities to develop needed knowledge and skills?
- What are the relationships between school factors and student outcomes, especially those related to test-based achievement?

The ETS team proposed a novel design for the NAEP instruments and introduced a number of methodological innovations that changed the face of large-scale assessments. The proposed design was properly titled "A New Design for a New Era" (Messick et al., 1983). The innovations and the subsequent refinements of this design have been reviewed by Mazzeo et al. (2006). With respect to instrument design, the assessment frameworks became more directly tied to the underlying constructs (Kirsch, 2003), enhancing construct validity.¹⁴ Among the methodological innovations were the use of balanced incomplete block (BIB) designs for the booklets administered to students (Mazzeo et al., 2006) and item response theory (IRT) in order to be able to report the results from multiple forms onto a common scale. Subsequently, ETS introduced latent regression models and Bayesian inference procedures to generate multiple imputations of the results in the form of "plausible values" that are now used to obtain population estimates and their corresponding measurement errors (Braun & von Davier, 2018; von Davier & Sinharay, 2014).

The novel NAEP design and methodology marked a key inflection point in the history of large-scale assessments. It set NAEP on a new trajectory that continues to this day. Moreover, these innovations were adapted by the different emerging ILSAs and, to this day, are at the core of the ongoing evolution of designs, methodologies, and platforms (see Mullis & Martin, 2019, for an extended case study). The coordination of these innovations and extensions led to an ever-rich body of information that ILSAs provided to policy makers, researchers, and other key stakeholders whose information needs have been evolving in tandem with changes in the social and economic landscape.

Indeed, starting in the 1990s, there was a growing appreciation of the significant connections between human capital and important outcomes both for individuals

and for the societies in which they live. In the United States, for example, the report by the Secretary's Commission on Achieving Necessary Skills (Kane et al., 1990) focused on identifying the skills needed for the country to enjoy a high-skills, high-wage, high-productivity economy. At the international level, the OECD (1992) noted that low literacy levels were a serious threat to economic performance and social cohesion. And the Definition and Selection of Competencies project provided a theoretical and conceptual foundation for a broad range of competencies that individuals would need to meet the changing demands of modern societies (Rychen & Salganik, 2001, 2003).

As a result, policy makers began to ask new questions focused on adult populations: How are educational attainment and cognitive skills related? How are literacy and numeracy skills related to health and well-being, as well as to participation and success in the labor force? What factors may contribute to the acquisition and decline of skills across age cohorts? How are literacy skills related to voting, trust in institutions, and other indices of social participation? (Kane et al., 1990; OECD, 1992). More recently, there have been organized efforts to develop taxonomies of 21st-century skills in an international context (e.g., Rychen & Salganik, 2001). Binkley et al. (2012) provided an extensive review. As noted earlier, this growing interest led to a series of international assessments focusing on adults, culminating in PIAAC.

Much like the 1983 design for NAEP that put large-scale assessments on a new trajectory, PIAAC marked the beginning of a significant cycle of innovation. As the first ILSA to be designed from the outset as a fully digitally-based assessment, PIAAC expanded what could be measured. It included, for example, technology-based tasks that more directly reflect the changing nature of how people access, use, and communicate information (Leu et al., 2021). In the workplace and in everyday life, it is increasingly important for adults to be able to navigate, critically analyze, and problem-solve in complex, data-intensive digital environments—and the PIAAC platform has made it possible to measure such skills. In addition, PIAAC introduced methodological innovations such as multistage, adaptive testing and more flexible routing for the background questionnaires that have improved the design and delivery of the survey, laying the foundations for future assessments (Kirsch et al., 2017). Other PIAAC innovations include the implementation of interactive stimuli and automated scoring of tasks administered in more than 30 languages, including character-based languages. Consequently, we regard PIAAC as marking another inflection point in the history of large-scale assessment surveys.

Countries now participating in international student and adult assessments represent the majority of the world's gross domestic product, with participation from low- and middle-income countries continuing to grow. That growth is fueled in large part by sponsoring organizations such as the World Bank, UNESCO, and the Inter-American Development Bank (Lockheed, 2013). These organizations regard international assessments as a cost-effective way to monitor the efficacy of their

educational investments in countries where national assessment systems, if they exist, are of variable and at times questionable quality. Their investments represent a growing recognition of the importance of monitoring learning and skill acquisition in support of economic and social development. This shift is also reflected in the United Nations' Sustainable Development Goals (specifically SDG 4 on quality education), which include learning targets and not just time spent in school as evidence of having achieved this goal.

There is a concomitant interest in enhancing the policy utility of ILSA data: Policy makers and other key stakeholders, including researchers representing a broad range of disciplines, are calling for these assessments to measure new and important cognitive domains to provide richer background and contextual information. The intent is to provide deeper understandings of how skills develop and how they relate to educational, social, and economic outcomes. These policy-driven questions provide the impetus for what we refer to as a virtuous spiral (Figure 20.2). Such questions lead to the formulation of new assessment frameworks that guide the development of new instruments. The desire to measure novel constructs such as collaborative problem-solving, as well as to expand the measurement of existing constructs (e.g., incorporating electronic texts in the literacy domain), drives further advances in assessment designs and statistical



FIGURE 20.2
The Virtuous Spiral

models that facilitate richer analyses and deeper interpretations of the data. These, in turn, elicit increased interest among a wider group of stakeholders, leading to further questions. The result is that ILSAs evolve along this spiral of increased relevance and policy utility.

THE MESSICK FRAMEWORK, EXTENDED

Inasmuch as the primary purpose of an ILSA is to provide credible evidence to inform policy makers and other stakeholders, Messick (1987) proposed a framework consisting of several design criteria that could offer guidance for the design, development, and implementation of large-scale assessments. These include comparability, interpretability, and relevance. An overall judgment of the potential for policy utility is then largely dependent on evaluating the strength of these criteria with due regard to purpose, context, and the population(s) of concern (see Kirsch & Braun, 2020, for an exposition of the original Messick framework).

Now, 35 years later, paper-based instruments have given way to digitally-based assessments that build on the past but draw on technological innovations in various domains. These include the use of digital platforms and new electronic tools, along with innovative workflows and processes. They also comprise advances in measurement science and improved methodologies to analyze complex data originating from multiple cultural and linguistic sources (more so than ever before). This point in time is an opportune moment to extend and refine Messick's three design criteria and to examine how they are impacted by technology, both directly and indirectly. This analysis leads to suggestions for innovations that could substantially enhance the policy utility of ILSAs.

Comparability refers to the degree to which the results obtained from a range of reporting groups, representing different cultural and linguistic contexts, have the same meaning in relation to the underlying constructs. In the context of an international assessment, achieving comparability is essential to policy utility. However, the demands are significantly greater than in a single national context because of the much greater heterogeneity that must be addressed. For example, sponsors and developers must make every effort to ensure that the samples of respondents in the different countries are approximately equivalent in the statistical quality of their representations of the corresponding target populations. The degree of success depends on such factors as the nature of the auxiliary information available for the survey design, the degree of cooperation among sampled units, and fidelity of implementation (Rust, 2014). These factors vary across countries, and weakness in any one of them undermines sample quality and, therefore, comparability. If approximate equivalence fails to hold, the country's results must be either reported separately or not at all.¹⁵

Further, the data generated by the cognitive instruments and background questionnaire must have equivalent meanings (measurement invariance) across cultures and languages (Dept et al., 2025; van de Vijver, 2018; von Davier & Sinharay, 2014). Accordingly, the survey instruments, including the background questionnaire, undergo an

iterative process of development, translation/adaptation, verification, and review (Ebbs & Wry, 2016; OECD, 2016, 2020). Following data collection, psychometric analyses reveal the extent to which the goal of measurement invariance has been achieved. If certain items appear to function differently in a particular country (or set of countries), appropriate adjustments are made to the psychometric models that are used to estimate item parameters and generate the reported scale scores for that country (Fishbein et al., 2020; OECD, 2016, 2020).¹⁶

Comparability also has a chronological dimension, namely, that cognitive scale scores, as well as composite scale scores derived from the background questionnaire, can be meaningfully related to the corresponding scores from earlier administrations. In the cognitive domain, this generally entails conducting scale linkage procedures that place new items from the current administration on the established scale. On occasion, a new scale is defined and the items from previous administrations are rescaled and placed on this new scale. In either case, the relevant procedures are well known, as are the checks on the validity of the linkage (Mazzeo & von Davier, 2014). With regard to the background questionnaire, the chief requirement is to retain the item set contributing to the composite scale or—if changes are absolutely necessary—to make only minimal alterations. Significant changes may disrupt the ability to continue an existing/earlier scale and require that a new scale be established. This problem arises because the number of items contributing to a scale is typically very small; hence, the loss of one or two items can have a substantial impact.

Interpretability depends, in large part, on the extent to which the instruments administered have been developed through a fully coherent process so that the reported scores can be given substantive or normative meanings that are credible, defensible, and accessible to a range of stakeholders.

The term *full coherence* signifies that each of the key operations of design, development, scoring, scaling, and reporting are not only appropriately linked to the intended measurement goals, but also functionally integrated through continuing collaboration among teams specializing in each of those operations. Thus, the term implies strong, evidence-based support for the desired interpretation(s) or, in other words, demonstrable construct validity (Messick, 1989).

Indeed, validation of the desired interpretations, whether of the cognitive scores or of the background scales, requires a thorough explication of the assumptions underlying the interpretations and an evaluation of the evidence supporting those assumptions (Kane, 2013). In this regard, Pepper (2020) argued that validity efforts in the ILSA context fall short of the guidelines in the 2014 testing standards (American Educational Research Association et al., 2014) and the requirements of the validity argument as articulated by Kane (2013). Pepper took as a case study the mathematics self-efficacy scales in PISA 2003 and PISA 2012, arguing, for example, that essential validity evidence that would be generated by (cognitive) response process analyses is entirely absent. Clearly, further work in this direction should become more of a priority for future assessments.

Relevance is the extent to which (a) the evidence elicited by the cognitive instruments and the background questionnaire is germane to current policy questions and decisions and (b) the assessment design yields results that can be analyzed in such a way as to address current priorities. Foundational to relevance is the designers' reliance on the assessment and questionnaire frameworks developed by international experts and reviewed by participating countries—with an eye on the most construct-relevant and policy-appropriate elements of each domain and construct that are measured. Of course, relevance is strongly dependent on both comparability and interpretability; that is, a deficiency in either one directly undermines the utility of the data in addressing the questions of interest. Judgments of relevance are made by the various stakeholders in each country, as well as by secondary analysts, and are contingent on the particular purposes at hand. The voluminous technical reports accompanying an ILSA provide useful descriptions of instruments and methods, as well as data displays to inform those judgments.

In addition to ensuring comparability and interpretability, ILSA designers adopt different strategies to enhance relevance. One example is modifying the operational definition of *legacy constructs* and extending the corresponding assessment frameworks. Often, new item types are introduced to target new or neglected facets of the construct. Recent reading literacy frameworks in PISA and PIAAC are good examples: The implementation of digitally-based assessments facilitates the introduction of electronic texts into the assessment. Such texts are an increasingly important source of obtaining and communicating information and, hence, deserve the emphasis they receive in the most recent assessment frameworks. For another example, TIMSS (which is now fully digital) has introduced problem-solving and inquiry tasks to simulate problems arising in laboratories and in the real world that require students to apply a combination of procedural skills and content knowledge. In the case of the background questionnaire, examples include the development of measures of certain aspects of learning contexts (or other background factors) that research indicates may be associated with the development of cognitive skills.

We believe that the improvements to be described here contribute to greater comparability, interpretability, and relevance, resulting in increased utility for policy makers and key stakeholders. Increased utility will continue to drive the salience and growth of these surveys that, in turn, will raise new questions, requiring further innovations on the part of developers and contractors. It is this productive dynamic that we have termed the *virtuous spiral*, depicted in Figure 20.2.

From an operational perspective, digital technologies provide the field with a powerful new infrastructure (platform) that both inspires and facilitates the development and refinement of new tools, processes, and workflows. Accordingly, the transition can impact all major phases of ILSAs: management, design, development and delivery, data handling, analysis, and data product generation. Although each of these phases exists with paper-based assessments, their development and implementation in

technology-based platforms now requires higher levels of coordination and integration to achieve the anticipated gains in efficiency and data quality.

TRANSITIONING TO DIGITALLY-BASED ASSESSMENTS

In a digitally-based, large-scale assessment, a well-designed platform makes it possible to introduce efficiencies that positively impact the development and conduct of the survey, including delivery, data capture, and processing operations. For example, the platform facilitates innovations in assessment design and delivery, allowing for embedded routing and branching capabilities that control how respondents move through the various components of the assessment. In particular, the implementation of multistage, adaptive testing models makes it possible to carry out more efficient measurement.

Platforms may also include a portal that functions as a centralized location for monitoring the workflows associated with key tasks related to the conduct of the survey, along with the interactions between contractors and each participating country. These enhancements serve to improve standardization of each stage of the work and allow for prompt interventions and adjustments, should they be necessary. They also can provide (a) item writers with capabilities for new item types and response modes to improve the measurement of legacy constructs or the introduction of new constructs and (b) translators with an improved infrastructure for translating and adapting assessment materials as well as receiving feedback from the verification process.

From the perspective of participating countries, the digital platform can not only support but also improve operational activities. The possibility of automatic data capture and scoring introduces efficiencies in terms of a reduced data entry and scoring burden along with more consistent scoring across countries and languages. Digital platforms also introduce new capabilities such as the capture of a full range of process data associated with respondent actions when interacting with assessment tasks, as well as accommodations that lead to greater accessibility for students and adults participating in the surveys. A full description of the potential efficiencies of a digital platform is beyond the scope of this chapter. This section, however, discusses the platform's role in a subset of key activities. Collectively, these enhancements can have significant impacts on the relevance, comparability, and interpretability of an ILSA, thereby increasing its overall utility to policy makers and key stakeholders.

Survey Instrument Development

The platform must possess the functionality to support item development and authoring for both the cognitive instruments and the background questionnaire in multiple languages and orthographic systems. Consequently, platform developers engage at the outset with domain experts and instrument developers to understand

the technical implications of (a) measuring new constructs, (b) revised and extended frameworks that broaden what can be measured, (c) delivering new item types, and (d) the approaches required to address a greater range of respondents' proficiencies due to the expansion in participating countries. At the same time, domain experts and instrument developers take into account the range of displays, item types, and interaction modes that the platform supports, resulting in some modifications of item types and/or response modalities. This collaboration often results in timely platform development and enhancements that contribute to the relevance and interpretability of what is being assessed.

Assessment designs for paper-based surveys are limited by operational constraints related to time and costs associated with assembling, checking, printing, shipping, and handling multiple paper forms. By contrast, digitally-based assessments, especially when combined with some form of adaptive testing (see the section "Survey Administration"), offer many advantages. First is the potential to eliminate errors associated with the assembly and distribution of multiple paper booklets both within and across languages. Second, the digital platform can accommodate designs involving complex routing algorithms that entail a very large number of virtual forms. In particular, it can assign an appropriate, predesigned form to each participant, precisely following the assessment design (a practice that is particularly germane to adaptive testing). Moreover, the platform can support both instrument administration in different formats and data capture for a range of response modalities. Finally, greater efficiency in the estimation of proficiency distributions, in combination with new item types, makes possible both broader construct coverage of existing domains and the assessment of new constructs, topics we address in the sections "Developing Coherent Cognitive Assessments" and "Background Questionnaires."

Developing Coherent Cognitive Assessments

Following a coherent process for assessment development enhances the interpretability of the findings. In this regard, a construct-centered or evidence-centered approach to assessment development is most helpful (Messick, 1994; Mislevy et al., 1999). Successful implementation yields a consensus on an operational definition of the target construct; thus, it provides assessment developers with a road map for the design and development of the tasks, as well as for the collection of evidence that can be used to represent performance in relation to the construct. This process yields a reporting scale that can be more appropriately interpreted.

The assessment framework for each domain contains specifications for item development: (a) the identification of key task characteristics to be varied singly and jointly, (b) the numbers and types of items (stimulus materials, response formats, and levels of difficulty) required to populate the item pool, (c) guidelines for instrument assembly, and (d) considerations related to scoring open-ended items (see Lennon & Kirsch, 2025, for an example). In advance of field testing, item quality is evaluated in terms of the items' links to the assessment framework and a judgment that

their formats, layouts, and content follow established design principles. Subsequent to the main assessment, as well as the scaling and analysis of the cognitive data, the assessment framework facilitates the creation of item maps and the validation of task characteristics. These are often used to characterize proficiencies at different points along each cognitive scale.

In an international context, development of cognitive assessments that yield scores that are valid and comparable requires that every effort be made to establish the items' cross-cultural, cross-linguistic, and cross-national validity. In that regard, the digital platform should facilitate (a) flexible content management, so that development files can be shared as needed; (b) item previewing, so that items can be examined in both source and target languages, along with the corresponding item layouts; and (c) consistency in item translation/adaptation across the item pool. Thus, the platform must support the active involvement of participating countries, with continuous coordination among geographically dispersed teams, comprising domain experts, test developers, psychometricians, platform developers, and graphic designers.

Background Questionnaires

ILSAs also yield information on a set of constructs that are measured through the background questionnaires. In addition to standard demographic information, the constructs targeted by the background questionnaires are based on frameworks developed through collaborative efforts among representatives of participating countries, international content experts, and questionnaire developers. One goal is to assess important characteristics or factors associated with students, teachers, and schools for surveys focused on in-school populations and important social, educational, and labor market factors for adult assessments. In addition to these context-related constructs, school-based surveys also target students' attitudes and behaviors along with family-related characteristics. Although these constructs are of interest in their own right, they are particularly valuable because of the insights that can be gleaned from their relationships with the cognitive outcomes, as well as the differences in those relationships across populations and through time.

Since the mid-1990s, background questionnaires have become more comprehensive in scope and their development more systematic. For example, rather than focusing on authoring individual items, the trend has been to develop sets of items whose responses can be combined into a scale that relates directly to a target construct, placing a new burden on item developers and the committees that advise them. Scale construction by means of IRT modeling is now accepted practice, with due attention to model fit, reliability, and validity (Martin et al., 2014).

As stakeholders have come to appreciate the value of the background questionnaire, it has gained in importance, thereby putting more pressure on improving the assessment of existing constructs while accommodating new indicators—all within strict time constraints for administration. The result is an ongoing tension between introducing these new indicators and maintaining trend for (at least some) existing indicators. This

tension provides the impetus for methodological developments, such as incomplete designs analogous to those employed for the cognitive instruments. PISA 2022 implemented a within-construct rotation design for the background questionnaire. However, implementing incomplete designs for the background questionnaire makes strong assumptions about conditional independence of context variables and has implications for the scaling of the cognitive domains that have not yet been fully resolved (von Davier, 2014).

Nonetheless, this dynamic provides yet another instance of the virtuous spiral introduced in Figure 20.2. As with the cognitive instruments, the platform plays an important role in supporting more complex designs and routing patterns often associated with current background questionnaires. These improvements support the relevance of the information that is gathered, as well as the quality and consistency of the information that is captured.

Translation/Adaptation/Verification of the Instruments

An essential step in establishing cross-national comparability is the translation, adaptation, and verification of the items in the cognitive instruments and the background questionnaire. This process is difficult, costly, and time-consuming. By supporting new workflows, the digital platform contributes to making the process both more accurate and efficient. As noted earlier, the platform accommodates the full range of languages used by participating countries (including right-to-left and ideographic languages) in addition to supporting the coordination of the work of test developers, linguists, and cross-cultural survey methodologists. This team produces informative translation and adaptation notes that both explain the underlying constructs and offer guidelines for use during the test translation and adaptation process. These improvements in this work process help to improve quality control, thus facilitating standardization and resulting in improved comparability.

In the case of PIAAC, for example, the item-by-item translation and adaptation guidelines (a) explain what the item is intended to measure; (b) specify which adaptations are mandatory, desirable, acceptable, or ruled out; (c) draw the translators' attention to terminology problems, translation traps, and patterns in response options; and (d), in the case of recurring elements or elements already present in trend materials, indicate how to access previous translations of these elements.¹⁷

For the cognitive items, the guidelines provide information on certain crucial assessment-related features such as literal matches (e.g., between stimuli and questions) that need to be maintained in the translated national versions, level of language difficulty, distractors, and so on.

A specially developed set of integrated tools in the platform makes it possible for these guidelines to appear in the translation tool when a translator on the national team (or reconciler, or verifier) processes a text segment. This is a technical innovation offering significant added value by streamlining processes, reducing the number of documents and tools required, and providing a translation environment that unites

all relevant information—thus enabling translators to better attend to key elements of the translation task at no additional cost for countries. We expect that forthcoming artificial intelligence–based tools will offer further enhancements and efficiencies to the translation/adaptation process.

Population Sampling Operations

The primary goal of an ILSA is to obtain credible and comparable estimates of population distributions of proficiencies in cognitive domains. The essential first step in sampling is a clear definition of the target population. Then a probability (random) sample of units is selected from the population, because randomization justifies using the machinery of probability theory to make inferences from sample characteristics to population characteristics. An added benefit is that with a well-conducted survey it is possible not only to obtain approximately unbiased estimates of population parameters (e.g., means, variances, percentiles), but also to quantify the uncertainty attached to those estimates (e.g., estimates of the standard errors of the parameter estimates).

Because of considerations related to cost and logistics, as well as reporting requirements, large-scale surveys almost never draw simple random samples of the units of interest (students or adults); rather, they employ complex sampling designs such as multistage cluster designs.¹⁸ An important design consideration is the trade-off between the amount of information derived from a sampling unit and the cost of obtaining that information. The key is that, at least in principle, there is a known probability used for the selection of each unit. These probabilities are used in the calculations leading to the estimates of population parameters and their variances.

The principles and techniques of population sampling are treated in many texts (e.g., Lohr, 2010). Rust (2014) presented a comprehensive treatment of considerations in ILSA sampling. TIMSS and PIRLS sample schools and students in such a way as to be able to estimate proficiency distributions for particular grades. Consequently, they typically employ multistage cluster samples where schools are first grouped into relatively homogeneous categories (termed *strata*) characterized by such factors as geographic region, political status, and school type, as well as characteristics related to mean school achievement. At the country level, the choice of characteristics to classify schools for the purpose of selection also depends on the information available, the interest in the reporting accuracy of the results at various subnational levels, and considerations of efficiency. Within strata, schools are selected with probability proportional to size.¹⁹ The clusters are randomly chosen classrooms within the schools providing instruction in one or both of the target grades. The assessment instrument is administered to all students within the selected classrooms. This approach is efficient, is minimally disruptive to the school, and facilitates estimating relationships between student achievement and their teacher or classroom characteristics. The final sampling plan represents the ideal plan adapted to the real-world constraints particular to each country.²⁰ ILSAs recognize that achieving a sampling plan with 100% coverage is not realistic. Typically, countries are not annotated if they have no more than 5% exclusions. At the school level, reasons

for exclusions include schools located in difficult-to-reach regions or particularly small schools; at the student level, reasons may be nonofficial language speakers and students with disabilities.

By contrast, PISA estimates proficiency distributions for a particular age cohort (i.e., 15-year-olds). Accordingly, its design must take into account not only that students of the same age may be enrolled in different grades, but also that those grades may be located in different system levels (i.e., lower secondary and upper secondary). Consequently, there may be curricular differences that, along with other factors (e.g., differential participation and coverage), can impact cross-national interpretability of the results.

The sample design for PISA 2018 also utilized stratification for school selection. As with the IEA studies, the choice of stratification factors depended on the country. Within each selected school, a random sample of students is obtained from the student population in the targeted age cohort. The number of students selected within each school varies from 35 to 42 depending on the assessment options chosen by the country (see OECD, 2020, chap. 4, for further information).²¹

PIAAC is a household survey. For some countries it employs a type of multistage, clustered area sampling with households as the ultimate sampling unit. Within a selected household, the instrument is administered to a randomly chosen adult whose age falls within the target range of 16 to 65. In other countries with registries of either households or individuals, sampling is conducted directly from the registry. Consequently, the actual implementation of the sampling procedure varies by country, depending on the structure and completeness of the sampling frame. The type of procedure employed also affects variance calculations (see OECD, 2016, for more detailed information).

As is the case with all fieldwork, the obtained sample in a school-based survey differs from the ideal sample in a number of ways. An especially challenging problem is the existence of schools that are not included in the sampling frame. For those schools in the sampling frame, selected schools may refuse to cooperate and selected students within schools may be absent, refuse to cooperate, or not fully respond to the cognitive instrument and/or the background questionnaire.²² In the case of household surveys, the sampling frame or the registry may be inaccurate or incomplete. Once households are selected and visited, there may be no one at home or the selected respondent may refuse to participate or only partially respond to the survey.

Inasmuch as the utility of cross-national surveys is critically dependent on the quality of the samples drawn, each ILSA has a set of procedures to address sample quality issues. These procedures are documented in the technical reports referenced above (see Rust, 2014, for general considerations about sample designs and procedures). Procedures may include replacement sampling for noncooperating units and technical adjustments for various types of nonresponse. Notwithstanding the use of such procedures, substantial differences across countries in response rates at the various stages of design implementation reduce the credibility of comparisons of proficiency distribution estimates as well as other targets of inference. ILSAs typically set thresholds for response rates; results for countries not meeting the threshold are reported separately

or not at all. Nonresponse bias analysis is also conducted to estimate the effects of these issues on the overall estimates.

When information about the population of interest is lacking or incomplete, ILSA teams face significant challenges in ensuring sample quality within each participating country.²³ This may be a particular concern with household surveys. Because many of these countries have limited experience and capacity in conducting large-scale surveys, they require both extra support and more careful monitoring. In some countries, information to develop adequate sampling plans for national samples may be insufficient. At the same time, ILSA participation offers country representatives intensive training and the opportunity to garner field experience in the selection of probability samples that can be employed for later in-country surveys as well as future ILSA administrations.

ILSA teams are always seeking ways to reduce survey error. A model for improving data quality is the total survey error (TSE) framework, originally developed by Hansen and explicated by Hansen et al. (1953). The TSE framework covers all types of errors that may arise in survey design, sample selection, data collection and processing, scaling and analysis, and creation of data products. The TSE framework makes a distinction between sampling and nonsampling errors. Sampling error results from variability in the estimates stemming from the selection of a random fraction of the target population. Nonsampling errors may be introduced at any phase of the survey process. Consequently, they must be taken into consideration throughout survey operations and management (including development and implementation) as well as data collection, handling, and analysis. Operationalizing the TSE framework requires striking a balance between enhancing data quality and operating within survey constraints (Biemer et al., 2017).

The transition from paper-based to digitally-based interviews and assessments has enabled many advancements in error control. One strategy is to establish processes that detect various sources of nonsampling error during data collection and to remedy them when possible. In this regard, the digital platform plays a key role through its capacity to collect and present what is referred to as *para-data* in a timely manner (Mohadjer & Edwards, 2018).

Para-data are the survey process data that are generated during data collection. For example, in carrying out the PIAAC household survey, the para-data may include the record of contact information, instrument timings, voice recording of interviews, geolocation of the interview, and interviewer work activities (hours and travel routes). They contribute to indicators of data quality, costs, and interviewer effectiveness. Because para-data can be quite voluminous, they are collected and summarized in a performance dashboard comprising a set of survey control charts and data graphs that monitor sample yield by interviewer and overall response rates at different levels of aggregation, highlighting unusual outcomes. With real-time transmission of information, the performance dashboard allows survey managers to react to operational challenges in a timely manner, facilitating what is sometimes termed an adaptive data collection strategy.

Survey Administration

In this section we review two modes of administration—paper and pencil and digitally based.

Paper-and-Pencil Administration

In preparation for this mode of administration, the item pool for a cognitive domain is organized into disjoint sets of items, termed *blocks* or *clusters*. For each ILSA, blocks are created according to specific criteria related to the number of items or item sets, content coverage, distribution of item difficulties, and timing. In a matrix sampling design, students are presented with a booklet that comprises two or more blocks, along with a background questionnaire. For example, in the case of booklets containing two cognitive blocks, the blocks are systematically paired into booklets. They are usually organized according to a BIB or partially balanced incomplete block (pBIB) design (Mazzeo et al., 2006; L. Rutkowski et al., 2014). In a full BIB design, each block is paired once with every other block.²⁴ This ensures that it is possible to compute covariances for all pairs of items, thereby facilitating placing estimated item parameters on a common scale (see the section “Scaling, Population Modeling, and Proficiency Estimation”) and using statistical analysis techniques that require the use of a complete covariance matrix, such as factor analytic techniques. Because of logistical constraints, it is not always possible to employ a full BIB design and designers resort to using pBIB designs, in which not every pair of blocks appears together in a booklet. Nonetheless, with modern psychometric theory, it is still possible to place all estimated item parameters on a common scale.

As one might expect, there are many variations on a theme. In a focused BIB design, each student receives blocks assessing a single subject (e.g., reading in PIRLS). In an unfocused BIB, students are assessed in two or more subjects (e.g., mathematics and science in TIMSS; reading, mathematics, and science in PISA). In the latter case, it is possible to compute correlations between cognitive domains. Of course, the intended interpretations of the results of the psychometric analyses depend on the assumption that the responses to each item (block) have been produced by randomly selected samples of students.²⁵ This assumption is met by various spiraling strategies for booklet distribution (Mazzeo et al., 2006). Detailed descriptions for each ILSA can be found in the corresponding technical reports.

To accommodate the increasing heterogeneity in country-level distributions of proficiency, ILSAs have adopted different strategies. For example, in PIRLS 2021, the proportional distribution of booklets by difficulty varies by country. Booklets are categorized as more or less difficult, depending on the distributions of item difficulties in the component blocks. Countries with (assumed) higher levels of proficiency receive a greater proportion of the more difficult booklets, while countries with (assumed) lower levels of proficiency receive a greater proportion of the less difficult booklets. By orchestrating a better overall match between the assessment and the country’s students, it is possible to obtain more accurate estimates of the overall proficiency distribution within the constraints imposed by paper-based administration.

Digitally-Based Administration

As noted earlier, the transition to digitally-based administration facilitates the introduction of new item types as well as new strategies for administration. With regard to the latter, it is certainly possible to organize the sets of items for administration in predetermined sequences of approximately equal difficulty (as is done in paper-and-pencil administration) and assign each set to randomly equivalent samples of respondents. Of greater interest is the introduction of adaptive testing. Adaptation can be done at the level of the individual item or of sets of items.²⁶ The latter is usually referred to as multistage, adaptive testing.

In multistage, adaptive testing, after a set of items has been administered and responses have been evaluated, an algorithm implements a decision rule that, for the next stage, routes the respondent to one of a number of possible item sets. The algorithm accounts for the student's performance on the previous item sets as well as content and other constraints. The intent is to improve the final estimation accuracy by better matching item difficulties to the current estimate of the respondent's proficiency, which is especially important when dealing with substantial heterogeneity in proficiencies within and among countries. In particular, adaptive testing reduces the chances that students will be exposed to items that they find extremely easy or extremely difficult.

Happily, it appears that adaptive testing induces greater respondent engagement, as indicated by lower rates of nonresponse or random responding. In conjunction with timing data that enable making distinctions between omitted items and items not reached, greater accuracy in proficiency estimates is achieved. At present, one difficulty is that in some ILSAs many item sets include one or more items that require human scoring. In that case, the decision rule must rely solely on responses to the machine-scored items. As a consequence, the selection of the next set of items is based only in the performance on these machine-scored items. This results in somewhat less improvement in efficiency than would be the case if the responses to all items were used.

With the large item pools typical in ILSAs, the number of item sets (or testlets) is correspondingly large, and consequently for multistage, adaptive testing, the number of different item paths (sometimes referred to as *virtual forms*) can run into the many thousands. Clearly, having such a large number of forms in a paper-and-pencil administration would be both infeasible and unaffordable. At the same time, the key innovation in matrix sampling has been retained, namely, that each respondent is only administered a small proportion of the full item pool.

Adaptive testing in an international context was first carried out in an individual-level test of adult literacy, setting the stage for later, more advanced implementations—first in PIAAC and subsequently in PISA. Yamamoto et al. (2018) discussed some advantages of multistage, adaptive testing over fixed form tests, especially in the ILSA context. They also note a number of challenges including (a) avoiding (to the extent possible) items that require human scoring in the decision-making process,²⁷ (b) ensuring that each virtual form meets construct representation requirements, (c) maintaining desired levels of item exposure control, and (d) employing appropriate analytic procedures. Indeed, implementing multistage, adaptive testing adds considerably to the complexity of the

assessment system. Yamamoto et al. provided a comprehensive discussion of these complexities and their resolution in the context of PIAAC and PISA.

Accommodations for Testing

Comparability of national samples is an essential element of ILSA utility. One potential threat to comparability is related to the assessment of individuals with disabilities, inasmuch as countries' practices differ in a number of ways: how they identify students with different types and degrees of disability, the nature of the accommodations typically employed during testing, and their protocols for exclusion from an ILSA. Attempting to harmonize these practices across participating countries is evidently infeasible.

On the one hand, the baseline strategy adopted by TIMSS and PIRLS is to review the accommodations proposed by the countries and to approve them, unless there is a clear threat to the validity of the assessment. For example, reading questions to a blind student may be acceptable in TIMSS but not in PIRLS. Clearly, allowing variation in accommodations represents a different level of standardization of assessment administration. On the other hand, the purpose of test accommodations is to allow students to demonstrate what they know and can do, but with only minimal changes to the target construct (Thurlow, 2014; see also Zwick and Rodriguez & Thurlow, both in this volume). To the extent that this is the case, and if students are afforded accommodations with which they are familiar, then a certain level of comparability is achieved—in the sense that each student has a fair opportunity to meet the challenges presented by the assessment. From an international perspective, this protocol prioritizes obtaining better estimates at the national level (for a country that includes these individuals in their target population) at the expense of some loss in cross-country comparability in estimates of proficiency distributions.

PISA does not employ accommodations in the main assessment, with two exceptions: (a) allowing for extra time and (b) administering a shortened (1-hour) version of the full assessment with a restricted set of response formats. Schools designate which of their selected students are eligible for one or the other of these accommodations, which are available for math and reading only. For PISA 2025, all items in the shortened assessment will be fully accessible, though not all item types will be represented in the assessment. The data from these items do not contribute to the estimation of the international IRT parameters, but they do so in the estimation of the plausible values (PVs) for the national distributions.

With the advent of digitally-based assessments, new policies and protocols are under development. For example, the OECD sponsored a small pilot study to investigate the feasibility of offering accommodations for students with a range of disabilities (Laitusis et al., 2018). The results were mixed with respect to practicality and efficacy, especially for more complex item types such as simulations. In any case, the report's findings and recommendations will inform subsequent phases of implementation research for PISA 2025, as well as for TIMSS and PIRLS as they transition to digitally-based assessments. It is expected that future administrations will explore the feasibility of implementing affordances such as varying font size and text to speech.

Overall, accessibility will be enhanced as new items are developed according to the principles of universal design and added to the item pools. At the same time, this approach is more feasible for some item types than for others. For the latter, special modifications will be required to enhance accessibility. Because of cost constraints, the number of such modified items will be limited; hence, they may be located in a small number of forms that can be administered when particular accommodations are requested.

In this regard, one other aspect of ILSA administration in school settings bears mentioning. The number of different devices that students are using day to day is growing. Current planning contemplates allowing students to bring their own devices, as long as they meet certain interoperability standards, as well as building a platform that will host those devices, provided they conform to a number of requirements. Carrying out this plan will be quite challenging. It remains to be seen to what extent the profusion of devices with different capabilities introduces construct-irrelevant variance and impacts score trends.

SCALING, POPULATION MODELING, AND PROFICIENCY ESTIMATION

IRT scaling, population modeling, and proficiency estimation are essential components of the ILSA workflow. The accuracy of the model parameter estimates, as well as the estimates of the precision of those estimates, constitutes the statistical foundations for generating the appropriate, individual-level proficiency distributions that are ultimately transformed into population-level results. Accurate estimates are essential to achieving comparability across all relevant reporting groups and, hence, valid interpretations of the results.

Introduction

It is customary to provide countries with standard, mandatory software that is used to manage, integrate, and validate their national data. Participating countries employ the software for data processing as well as to run checks to ensure, as much as possible, that the within-country data capture and integration accurately reflect the values given by the sampled persons and/or the interviewers.

In particular, the software is able to produce a series of reports that give an overview of the data quality and field operations progress. Each country is required to generate and review these reports during field operations and again prior to submitting the database to the contractor for international processing. Along with the national database submission, each participating country submits supporting data documentation that provides the ILSA team with detailed information regarding issues or technical difficulties in the administration of the assessment.

For each country, the input to this segment of the workflow is a set of validated data files. The output is a set of files containing, among other things, the estimated parameters

of the score distributions for the overall population and for designated subgroups. By this point in ILSA history, the statistical and psychometric procedures have evolved considerably in complexity from those employed in the early years of NAEP. Paraphrasing von Davier and Sinharay (2014), current methods may be viewed as sophisticated imputation approaches that combine the several advantages of IRT with an explanatory approach based on collateral information to produce accurate subgroup results—despite the relatively short testing time and the sparseness of the data. The use of collateral information is especially important when participants are asked to respond to multiple domains within the same limited time. Further details can be found in von Davier and Sinharay (2014), as well as in the technical reports published by each ILSA. Analysis teams also expend considerable effort in checking the intermediate results at each stage of the process to ensure accuracy and comparability of the results while accounting for possible country-by-item interactions. Moreover, these procedures are continually refined to improve efficiency and to accommodate new demands from sponsors and participating countries.

For simplicity, the discussion that follows assumes a single latent trait for the domain of interest.²⁸ With the primary goal of producing unbiased estimates of score distributions for populations and subpopulations, ILSA assessment designers are faced with the formidable challenge of meeting the requirement of full construct representation under the constraint of limited testing time. For paper-and-pencil administration, various types of matrix sampling are employed (see the section “Survey Administration”). Such plans are characterized as “missing by design,” and the use of matrix sampling has implications for subsequent analyses. In practice, a limited number of forms are developed and the design can be represented by a matrix that displays the fraction of the full item pool that is contained in each form. Forms are distributed in such a way that each form is administered to random sets of respondents of approximately equal size.

For digitally-based assessments and, especially, with the advent of adaptive testing algorithms, the matrix designs have become increasingly complex. As is the case with paper-and-pencil administration, respondents respond to a relatively small fraction of the total item pool. However, on average, the forms administered to a group of respondents will be better matched to the group’s proficiency distribution. The desired group-level estimates can be obtained through the use of advanced statistical methods—assuming that each item or item set has been administered to a proper probability sample of respondents, that the sample sizes are sufficiently large, and that there are satisfactory linkages across forms.

The analysis proceeds in four stages:

- Stage 1, item calibration: estimation of the parameters of the item response model
- Stage 2, population modeling, including latent regression analysis
- Stage 3, generation of plausible values
- Stage 4, linking plausible values to the established reporting scale

Item Calibration

All ILSAs make use of IRT to calibrate the items in the item pool. Calibration involves estimating the item parameters that characterize the probabilistic relationship between test-taker proficiency (i.e., their location on the scale corresponding to the latent trait) and performance on the item. The items may be scored dichotomously (two possible scores) or polytomously (three or more possible scores). The psychometric models employed differ across ILSAs: TIMSS and PIRLS use the three-parameter logistic IRT model for multiple-choice items, while PISA uses the two-parameter logistic IRT model. The general partial credit model is used by all ILSAs for calibrating all other items. (For a discussion of the models used in IEA assessments, consult von Davier et al., 2020).

Field trial instruments incorporate designed collections of new items or new item sets, along with a set of previously administered items that will be used to estimate trends. The total pool of items is administered to large convenience samples to estimate the measurement invariance of the trend items, as well as to obtain preliminary item parameter estimates for the new items. These analyses are used to furnish initial estimates of how well the trend items and new items function, both within and across participating countries with regard to comparability and overall quality.

If the items are selected for operational use, then the response data obtained from the main data collection are used to update the item parameter estimates for the new items and to evaluate the degree of comparability across countries. Operationally, if an item appears to have substantially different psychometric properties for a particular country, it is calibrated separately for that country.²⁹

With regard to the background questionnaire, item sets addressing constructs such as attitudes, dispositions, and behaviors are used to generate scales for secondary analysis. The process is also carried out by employing IRT models. The models take account of the fact that most of these items use ordered response categories. Both PIRLS 2016 and TIMSS 2019 have used the Rasch partial credit model to conduct item calibration employing the full data sample, with each country contributing equally to the analysis. Subsequently, the reliability of the derived scales is evaluated and the scales are subjected to validation procedures, such as tests of unidimensionality.

For both the cognitive instruments and the background questionnaire, the field tests and main administrations are essential to the maintenance and replenishment of the item pools. The transition to digitally-based assessments, particularly with the introduction of multistage, adaptive testing, necessitates an increase in the sizes of item pools for the cognitive domains, with the demands for piloting and calibrating new items becoming correspondingly greater. Increases in the item pools are also driven by the need to accommodate greater heterogeneity among participants with respect to proficiencies in the focal domains. Further, with the concomitant exponential increase in the number of (virtual) forms, more complex assessment designs are needed to obtain sufficient information for calibration and evaluation of differential item functioning (Yamamoto et al., 2018). Such designs, made feasible by the

digitally-based assessment platform, evidently contribute to improved construct representation.

It bears mentioning that both PIRLS and TIMSS use country-level adaptive designs (and sometimes benchmarking population adaptive). Such designs employ rotations of harder/easier booklets with the proportions based on prior information from a previous cycle. The goal is to improve estimation, particularly at the low end of the proficiency scale.³⁰

Population Modeling

The background questionnaire incorporates a set of questions related to constructs believed to be associated with (cognitive) proficiencies. These include both background characteristics and contextual factors. The second stage of analysis makes use of these additional data to obtain more accurate estimates of proficiency distributions. Specifically, this stage involves estimating the parameters of a so-called latent regression model (Mislevy, 1991; Mislevy et al., 1992; von Davier, Khorramdel, et al., 2019; von Davier & Sinharay, 2014) in which an individual's unobserved proficiency is regressed on a suite of variables derived from their responses to the background questionnaire. This model is usually referred to as a *population model*. In practice, the full suite of background and contextual variables is replaced by a smaller set of principal components that account for approximately 70% to 80% of the observed variance.³¹ The reduction in the number of explanatory variables yields more stable estimates of the regression coefficients and the variance–covariance matrix of the model, but with negligible reduction in accuracy (Oranje & Ye, 2014; Thomas, 2002; Wetzel et al., 2015). It is also customary to assume that the residuals about the regression plane follow an approximately normal distribution. The entire system is placed in a Bayesian framework with the introduction of a unit normal prior for each individual. The prior is successively updated with each full cycle of the expectation–maximization algorithm that is employed in the estimation procedure.³²

Concatenating the IRT model and latent regression model yields the marginal probability distribution for the observed item responses, given the transformed set of background variables, the estimated item parameters, and the parameters of the latent regression. Multiplying the marginal distributions across respondents results in a likelihood function. The corresponding log-likelihood function is then maximized with respect to the parameters of the latent regression, while the parameters of the IRT model are held fixed at the parameter values obtained in the first stage of the analysis (to impose comparability constraints across participating countries).³³

We note in passing that maximization of the log-likelihood function is a very difficult exercise in numerical analysis and considerable effort has been expended in devising different approaches to carrying out the computations (von Davier & Sinharay, 2014). The output is an estimate of the conditional distribution of the maximum likelihood estimate (MLE) of the vector of regression coefficients, as well as the MLE of the variance of the distribution of an individual latent trait. The latter is held fixed at its estimated value in all further computations.

Plausible Values

As noted earlier, the methodology employed by large-scale assessments is not intended to yield point estimates of ability at the individual level but, rather, estimates of group-level distributions of ability. This task is accomplished by producing a collection of parameter estimates that are used to generate a set of imputed values for each individual. In this context, these imputed values are PVs. Collectively, the PVs associated with an individual make manifest the available information regarding their proficiency in the domain. The average of the PVs is an estimate of the expected score were someone with their vector of principal components to be administered the full item pool. The variability among the PV is used to estimate the total measurement uncertainty associated with an estimate of a group-level parameter.

To obtain a single set of PVs (one for each respondent in the sample), it is necessary to approximate the proficiency distribution for each respondent. Since that distribution is assumed to be normal, it is only necessary to estimate its mean and variance. An estimate of the variance was obtained in the previous stage. The mean depends on the covariate values associated with the respondent and the (estimated) vector of regression coefficients of the latent regression model. That estimated vector has a posterior distribution that is approximately multivariate normal, with mean equal to its MLE and a variance–covariance matrix equal to the estimated variance–covariance of the MLE.

A random draw from that posterior distribution produces a vector of regression coefficients, enabling full specification of the proficiency distribution for each respondent. A single independent, random draw from that proficiency distribution yields 1 PV for the respondent. These steps are carried out for each respondent to generate a full set of PVs for the sample. The entire process is then repeated N times to produce N full sets of PVs. The number N of imputations was decided by most ILSAs following recommendations by Little and Rubin (1987) for multiple imputations. Early choices of 5 or 10 were limited by computing power, data storage, and time constraints. As computing power increases, 20 or more sets of imputations will be common.

Linking

An important use of ILSA data is tracking changes in skill distributions over time at the national and subnational levels. Such comparability over time requires linking the proficiency scale of the current assessment to the scale established in prior assessments. Linking is accomplished by embedding so-called trend items in the survey instrument. Trend items are those items employed in the current assessment that were also present in earlier assessments. Roughly speaking, by comparing performance across administrations on these items, appropriate scale transformations can be estimated and applied. However, actually carrying out the procedure involves a range of technical issues related to both design and analysis (Mazzeo & von Davier, 2014). Note that trend items contribute to proficiency estimation just as nontrend items do. In this regard, a PISA study (von Davier, Yamamoto, et al., 2019) deserves mention. In this study, PISA data from 2000 to 2012 (comprising more than 2 million students) were reanalyzed in a single

linking model to better understand aspects of linking, as well as issues related to model fit. One consequence was improved procedures that maximized the use of trend information over more than two cycles.

Although every ILSA adopts a somewhat different approach to scale maintenance, each faces similar challenges. In particular, designers must consider such factors as the degree of construct representation exhibited by the trend items, the number of trend items, their colocation in blocks, and block placement. Further information can be found in Mazzeo and von Davier (2014) and in the technical reports associated with each ILSA.

Ideally, successive assessments (and the corresponding item pools) are constructed according to the specifications of a common assessment framework. The set of trend items should be both fully representative of that framework and sufficiently large to support stable estimates of the linear regression parameters that transform the current scale into the earlier one. Consequently, it is not unusual for the trend items in paper-and-pencil administration to constitute approximately 50% of the item pool.

For trend items to function as intended, and to minimize the (unwanted) contributions of sources of variation unrelated to the focal domain, the context for each item is held constant to the extent possible. In the case of individual items, their location within a block, as well as the psychometric characteristics and speededness of the other items in the block, should be controlled. An alternative strategy is to employ intact item sets or item blocks from earlier assessments. Although this controls the local context, it may make full construct representation more difficult to achieve. In paper-and-pencil administration, block position within a booklet must also be considered. Further complications arise when considering the different designs employed for administering a single content domain to a student (e.g., PIRLS) or for administering two or more content domains (e.g., TIMSS, PISA). For example, designs for the latter situation possess some advantages from a measurement perspective because they make it possible to take advantage of the correlation of proficiencies between domains. However, they make keeping the within-instrument context for trend items constant more difficult because it is now necessary to also consider the paired domain as part of the context.

Finally, the human scoring of constructed response items must be strictly comparable across administrations. Accordingly, the training of scorers for the current administration must replicate the training conducted in the previous administration. Close monitoring of scorer behavior is essential. One approach is to seed responses with known scores from previous administrations into the current workflow and to compare the scores assigned at the two time points. Systematic discrepancies trigger retraining and rescoring and, occasionally, eliminating the item for the discrepant group.

As is always the case with IRT, there is a fundamental indeterminacy in fixing the latent trait scale after calibration. Either by comparing the item statistics on the trend items in adjacent administrations or by concurrently recalibrating the items in the two administrations, it is possible to resolve that indeterminacy in the newer administration

so that item parameters are placed on the scale established in the earlier administration. A linear transformation then places the performances in the current administration on the reporting scale established at the outset of the ILSA. This process enables valid comparisons of score distributions across administrations.

Notably, there is an underlying tension between scale maintenance and the desire to expand the assessment framework and/or introduce new item types. Similar issues of scale maintenance arise when there is interest in making comparisons across cycles for those scales developed from the background questionnaire. For further details, as well as descriptions of the linking strategies employed by the paper-and-pencil administration associated with the different ILSAs, see Mazzeo and von Davier (2014). It is also typical that each ILSA administration poses unique challenges. For example, TIMSS 2019 saw the advent of eTIMSS, which was administered in about half of participating countries. This new administration necessitated a “bridge study” to link eTIMSS to the standard paper version of TIMSS. Further, a number of less difficult blocks of fourth-grade math items were also introduced with the goal of improving measurement accuracy for lower performing countries. Finally, eTIMSS included items (problem-solving and inquiry tasks) that took advantage of the affordances of digital administration but had no counterparts in the paper version. Carrying out the procedures of scaling and linking in this setting was considerably more complicated than in past administrations. For details, consult Martin et al. (2020, chaps. 11 and 12).

Although all ILSAs are transitioning from paper-and-pencil administration to digitally-based assessments, paper-and-pencil administration versions are likely to be required for some time to come: For some countries, the technical infrastructure is not sufficient to support digitally-based assessments, while in other countries with the appropriate infrastructure, some numbers of respondents will not have sufficient familiarity with technology. Consequently, as noted previously in the case of TIMSS 2019, estimating mode effects (i.e., the impact of administration mode on item parameters and the implications for secondary analyses) is an essential step in effecting the transition (Fishbein et al., 2018). Estimation of mode effects and making appropriate adjustments to place scores from the two modes on the same scale can be technically challenging (von Davier, Khorramdel, et al., 2019; von Davier, Yamamoto, et al., 2019).³⁴

As the operational aspects of digitally-based assessments have become more routine, capacity (and incentive) to introduce novel item types to enhance construct representation has grown. These have included items with new response formats (e.g., ranking, multiple response, drag and drop), as well as various types of interactive items such as simulations that are now used in the assessment of science. In each case, suitable psychometric models must be proposed and the item parameters must be properly calibrated. As the divergence between the paper-and-pencil administration and digitally-based assessment instruments increases, the challenge of placing the scores on the same scale will become more formidable and the results more dependent on model assumptions (see the section “Threats to Relevance” for further discussion).

Inasmuch as digitally-based assessments, especially with adaptive testing, can accommodate larger item pools, trend items may constitute only about 40% of the item pool. In the case of an adaptive assessment, it is critical that all the virtual forms administered contain sufficient numbers of trend items.³⁵ Thus, the challenge of successfully implementing a linking strategy is somewhat greater with this mode of assessment. Down the road, as the digitally-based assessment is administered with a greater range of delivery devices, estimating “device effects” will assume greater importance. Depending on the empirical findings, some restrictions on acceptable devices may have to be instituted.

Measurement Invariance

Issues related to comparability have already been mentioned in the sections “The Messick Framework, Extended” and “Population Sampling Operations.” The latter section addressed the procedures developed to obtain statistically equivalent probability samples across participating countries. Equally important is ensuring that the cognitive scores and the responses to the background questionnaire have comparable meaning across participating countries as well. The terms *measurement invariance* or *cross-cultural equivalence* refer to this notion of comparability of meaning.

Measurement invariance is a term that refers to the degree to which the underlying construct retains the same meaning across different settings. For ILSA, the settings are the different participating countries. In the ILSA context, the requirements for strict measurement invariance are very difficult to satisfy. Indeed, a persistent technical problem is the lack of agreement on how to quantify deviations from measurement invariance at the different levels of stringency. One approach in common use is multigroup confirmatory factor analysis. Although a number of fit statistics are available for such models, as well as for recent refinements, their distributional properties in the context of ILSA data are a subject of ongoing research (Byrne & van de Vijver, 2017; L. Rutkowski & Svetina, 2017; van de Vijver et al., 2019). Van de Vijver (2018) reviewed the recent literature and noted that Bayesian approaches (e.g., Carlin & Louis, 2000) may prove useful in this context. In the interim, psychometricians will continue to employ “rules of thumb” with an admixture of intensive data scrutiny. In the cognitive domain, current procedures, such as freeing the globally estimated IRT parameters for those items judged to have strong interactions with countries/languages, may not be sufficient.³⁶

Van de Vijver (2018) proposed a tripartite framework for evaluating and addressing threats to cross-cultural equivalence: construct bias, method bias, and item bias. In the context of ILSAs, he asserted that with regard to the cognitive domains and, particularly, the scales derived from responses to the background questionnaire, method bias is perhaps the greatest threat to comparability. He cited such issues as variations in response styles, familiarity with the types of questions, and even poor reading skills in school-based ILSAs. In the case of PIAAC, because of the nature of the administration, individual differences in listening skills and aural memory are also potential sources of method bias.

In the case of sample surveys, a number of indicators can quantify the degree of departure from the ideal. Similarly, with regard to measurement invariance, there is a tripartite framework for evaluating the level of measurement invariance achieved (van de Vijver et al., 2019). In order of increasing stringency, the levels are configural, metric, and scalar. There are statistical procedures that can be used to distinguish among the levels and, hence, suggest limitations on the interpretability of the comparisons among countries. (Consult van de Vijver et al., 2019, and the references therein for further details.)

Notably, van de Vijver et al. (2019) also discussed steps that are taken to enhance measurement invariance, particularly with regard to the background questionnaire, where linguistic and cultural differences play an important role. They emphasized the importance of developing consensus-based frameworks (at different levels of generality) to guide the overall design and item development. There is considerable overlap with the presentation in the section “Transitioning to Digitally-Based Assessments.” Item-by-country interactions are addressed by Glas and Jehangir (2013) and von Davier and Bezirhan (2022). Further discussion of threats to comparability is contained in the section “Threats to Relevance.”³⁷

Generating Results

With the reporting scale established, it is possible to obtain estimates of the parameters (e.g., means, variances, and percentiles) of the proficiency distributions of the reporting groups of interest (e.g., gender, socioeconomic status, immigration status). Parameter estimates are obtained by aggregating the PVs of the individuals comprising the reporting group. Typically, an interim (weighted) estimate of the parameter is calculated for each set of PVs, and the final reported estimate is a simple average of the interim estimates.

Variance estimation is not as straightforward inasmuch as the estimated variance of a group-level estimate has two components: one due to sampling and the other due to measurement error. The former reflects the fact that the respondents are a probability sample from the population (group) of interest and that, in a full replication, a different sample would have been obtained. The estimate of this variance component is derived using standard methods for sample surveys with proper attention to the nature of the sampling frame and the corresponding sampling weights. Again, different ILSAs use different estimation strategies (Rust, 2014; von Davier & Sinharay, 2014). The second component reflects the measurement error in the estimation of proficiencies. For a specific group-level parameter, this component is proportional to the average of the squared differences between the interim estimates and the final reported estimate.

Building on the process for instrument design and development, in conjunction with patterns of item responses, it is possible to describe, in substantive terms, the differences among performances at various locations along the scale. Such descriptions contribute to greater interpretability, thereby enhancing the utility of the assessment results to policy makers and other stakeholders (Kirsch, 2001; Mosenthal &

Kirsch, 1991). These methods have been applied to the cognitive scales constructed for PISA and PIAAC.

A somewhat different approach, based on the behavioral anchoring methods first developed for NAEP, is employed by PIRLS and TIMSS.³⁸ With this approach, specific benchmarks along the reporting scale are selected and psychometric analyses are conducted to select items that discriminate between adjacent benchmarks (Martin et al., 2020, chap. 15). Identification of commonalities among the items at each benchmark are then used to provide substantively grounded descriptions of performance at the benchmarks.

Design Issues

In ILSAs, it is commonly the case that multiple proficiencies are estimated or that a single proficiency has been decomposed into some number of facets. In that case, multivariate versions of the latent regression model can be implemented. The overall plan is the same, although the mathematical notation becomes more unwieldy and, more important, the calculations become substantially more burdensome (von Davier & Sinharay, 2014). The benefit, however, is that the analysis can take advantage of the correlations among proficiencies (or facets) to yield more accurate estimates overall.

However, this situation can be a double-edged sword. For example, a difficulty arises if multiple skill domains are assessed but are not evenly represented in the assessments administered to individuals. Suppose that each individual is only tested on a subset of the full set of skill domains. In that case, the appropriateness of the PV obtained for the skill that was not tested is strongly dependent on the correctness of the population (statistical) model. PISA is a case in point. For example, the PISA database contains reading PVs for students who received a test form with mathematics and science tasks—but no reading tasks. For these students, generating the PVs for reading depends on the correlation of reading proficiency with mathematics proficiency, science proficiency, and background variables based on data from those students who received a test form that combined reading either with mathematics or with science.³⁹

Consideration of this issue highlights the difference between tests that provide individual-level scores and tests that provide only group-level scores. In the case of the former, fully model-based scores would generally be unacceptable. In the case of the latter, because of the random allocation of domain question sets across the population sample, most of the individuals comprising the group would have been exposed to items from each domain. For example, in PISA 2018, reading was the main domain. Accordingly, all students took some reading items, approximately 54% took some math items, and approximately 54% took some science items. Consequently, the estimated group mean score will largely reflect the contributions of those data. However, the estimated standard error associated with that estimated group mean score will reflect the increased measurement error due to the cognitive data missing from one third of the group.

Supplementary Considerations

As noted earlier, an important use of ILSA data is the estimation of relationships between proficiencies and background data or contextual factors. By including the corresponding variables in the latent regression (or at least their proxies embedded in the principal components), approximately unbiased estimates of those relationships are obtainable if the model is correctly specified (Mislevy, 1991). Achieving this goal is an important aspect of the rationale for including all available information in the latent regression model.

At the same time, there is continuing pressure to include a larger number of scales in the background questionnaire. One suggestion is to adopt rotated (incomplete) designs analogous to those used in the cognitive instruments (i.e., between-construct rotation). However, unlike the situation for the cognitive scales where the assumption of an underlying latent trait induces a set of useful conditional independence relationships, such an assumption is not feasible for the background questionnaire. Moreover, it is not clear how to modify the latent regression model to accommodate such designs without making strong assumptions of conditional independence, whose failure would seriously compromise the quality of the estimates reported (von Davier, 2014). Nonetheless, efforts are underway to use within-construct rotations for some background questionnaire indices with a large enough set of constituent items.

An additional consideration arises from the fact that, in the context of a digitally-based assessment, it is possible to collect process data, and the log files that are generated can record every keystroke. The sheer amount of such data is overwhelming; research is being conducted on how such data might be compressed into a small number of indicators that could be incorporated into expanded latent regression models to improve estimation accuracy (Bergner & von Davier, 2019; Kroehne et al., 2016; von Davier, Khorramdel, et al., 2019). Another approach would be to use such indicators to assess the quality of the data before item calibration. For example, individuals whose log files suggest aberrant response processes or patterns could be removed. Models for capturing the speed/accuracy trade-off may prove useful here, as well as in secondary analyses (van der Linden, 2007; Yamamoto & Lennon, 2018). See the sections “Impact” and “Technical Challenges” for further discussion.

DATA PREPARATION, DATA PRODUCTS, DATA ANALYSIS TOOLS, AND RELATED ACTIVITIES

At the conclusion of the analysis stage, each participating country receives a set of databases specific to that country. Ultimately, an international database is constructed, combining certain country-specific databases. The sponsoring organizations, countries, researchers, and other key stakeholders employ the international database for a variety of purposes. ILSA contractors, in conjunction with the sponsoring organization and the participating countries, carry out a number of procedures leading to the dissemination

of the data that have been produced. They also engage in activities that promote the use of the data for a variety of purposes. This section briefly describes five key areas of activity: data preparation, data analysis and reporting tools, data products, data security and confidentiality, and data quality.

Data Preparation

Internal Data Processing

The main objective of this stage is to ensure that the data files adhere to international formats, that the data accurately reflect the information collected by each country, and that the different survey files can be linked appropriately. Upon receipt of the national databases, the contractor archives the data to a preliminary international database. The initial review of the national databases includes the evaluation of the supporting documentation and the required consistency among data sets. Each country is consulted during this step so as to address any issues or concerns.

Once the data files have been made consistent with the international database structure, as specified in the international codebooks, a set of programs is applied to conduct “data cleaning.” This process consists of three activities:

- checks related to the identification codes associated with the variables;
- checks related to linkage of variables within a data set and records between data sets, with particular attention to any discrepancies between the observed and expected data patterns; and
- checks related to cognitive and background variables.

Further quality reviews of the submitted data sets are conducted according to the technical standards and guidelines that are associated with each ILSA. These occur at multiple stages of the data preparation workflow. Typically, the contractor supports countries in issue resolution and uses an automated reporting mechanism to communicate with countries regarding the status of data submissions, quality review, and any data quality issues encountered. Once data cleaning and restructuring are completed, the files that include sampling information and the sampling weights are merged into each national database.

Creation of Derived Variables

After the international variables have been harmonized and all data validated, a set of derived variables is computed and added to the data set. A derived variable is typically constructed by combining two or more variables, resulting in a new variable or a scale score. These variables are fully documented, as required for further analysis and reporting activities. Derived variables undergo checks using state-of-the-art quality control methods.

Simultaneous with the creation of the derived variables, the cases in the data file are inspected to ensure that they are appropriate for inclusion in the data file. As a part of this evaluation, case sampling weights are computed and linked to the cases.⁴⁰ This

augmented data file is merged into the preliminary international database to yield a fully integrated international database (master database). Extracts from this database are then sent back to each country for its own use and forwarded to those responsible for data analysis.

Analysis and Reporting

Preparation of Databases

To support the export of data products meeting the needs of various users, different data versions can be generated. These usually comprise three files:

1. The national file submitted by the country and processed by the sponsoring organization. This file has all the data collected within a country, including additional subpopulations and variables that were part of the national data collection activities and samples. These files are typically only available from the national center and only to the sponsoring organization and the contractors.
2. A restricted user file with all the variables and cases intended for use in international comparisons. This file is typically used for secondary analyses. It is published by the sponsoring organization and usually available provided assurances of confidentiality are made by the user.
3. A public user file (PUF) contains all the variables and cases intended for use in international comparisons but might have some variables or values for them suppressed or masked. The intention of suppressing or masking values is to prevent the identification of participating individual or subgroups (i.e., schools, neighborhoods). These PUFs are posted on the Internet and can be accessed and used with no restrictions.

Once these databases are completed and have undergone preliminary quality control, standard programs are applied to generate multiple tables that will ultimately appear in technical reports and other documents associated with the ILSA.

Preparation of Public Use Files

Following the initial data-cleaning process, an iterative process of data review and correction takes place—initially among the contractors and later involving the participating countries as well as the sponsoring organization. By default, the PUFs maintain all international variables approved for release. They only include those records that fulfill criteria set for cases to be used in weighting and analysis.

Some procedures are specific to the sponsoring organization. For example, for OECD ILSAs, any and all national variables identified by a country for deletion are dropped. In addition, all international variables earmarked for suppression by a country are blanked (i.e., set to the appropriate missing value for all cases). By contrast, for TIMSS and PIRLS (sponsored by IEA), national variables are typically not included in the international database, but are made available to the nation asking these additional questions so the data can be merged. However, the international source versions

of context questionnaires can include items specifically created for that purpose. As an example, a small number of international variables can be provided for countries to include additional home possessions variables that could contribute to an indicator of socioeconomic status. These variables are labeled “country specific” in the database, and their English back translations are listed in a National Adaptations Database included with the user guide. Any context variables earmarked for suppression by a country are blanked. However, cognitive items may be first reviewed in consultation with the country. All such cases are documented in the user guide or in the technical report.

Each country’s database is included in the following types of electronic files:

- Data files in SAS and SPSS formats, with separate files for each respondent (student, teacher, principal, parent). Other formats are available for different ILSAs. This file (or multiple data files containing appropriate linking information and instructions) will contain all the variables approved for release to the public.
- For PISA and PIAAC, a data codebook document file in either PDF or EXCEL format. This two-part document contains a brief listing of the name, position, format, and description of each data variable, plus an expanded listing for the discrete variables that have coded information, including their code values, descriptors, and (unweighted) frequencies within each data set.
- For TIMSS and PIRLS, a codebook that provides a list of variables with the corresponding labels and formats. Additional information about variables is provided in the user guide or in the technical report. Finally, data almanacs contain frequency distributions and summary statistics for all variables.
- For TIMSS and PIRLS, a software package (IDB analyzer) is available for R, SPSS, and SAS versions. It is updated to include special routines to load and analyze data from IEA assessments.⁴¹
- Documentation that describes the contents and structures of the data files, explains how to employ the other resource files in the product, and provides the technical information needed to support users in conducting analyses of the data. This includes the names and uses of the key variables in each data set, as well as indicators of sample quality (e.g., response rates).

Preparation of Data Tables, Compendia, and Reporting Tables

SUMMARY TABLES FOR COGNITIVE ITEMS AND BACKGROUND VARIABLES One set of tables contains weighted summary statistics for each participating country on each cognitive item and on each variable in the background questionnaires. For the latter, summary tables display the international averages for each variable, with each country weighted equally. For each variable, the summary tables display the question that was asked, the location in the corresponding questionnaire, and the variable name in the data file. Note that the table format depends on whether the data are categorical or continuous.

The cognitive data summary tables (sometimes referred to as item analysis tables) contain summary information about the responses to the cognitive items. For each country, they contain weighted summary statistics, including variable identification, sample size, number of valid cases, weighted percentages of individuals corresponding to each valid response, weighted percentages of individuals who did not select any of the valid response options, and the average score on each of the cognitive domains. Standard errors are also included, where applicable. The item analysis tables are used by both the team and countries for further quality control, verifying data structure accuracy, and validation purposes.

Unweighted item analysis results are also generated because they are particularly useful in verifying the accuracy of the data structure. When necessary, for internal quality control, the team may produce separate analyses to compare item statistics for respondent groups taking different modes of assessment (for example, paper-and-pencil administration and digitally-based assessments) or for respondent groups determined by selected background variable categories.

COMPENDIA Using the international restricted user file and PUFs as the source data, compendia are sets of tables that provide the distribution of students or adults according to the variables collected by the questionnaires. The public version is essentially a redacted version of the summary data tables. The purpose of the public compendia is to support PUF users so they can better appreciate the contents of the PUF and, importantly, use the compendia results to verify that they are performing PUF analyses correctly. For confidentiality reasons, some countries may decide to alter their data or remove respondent records from their PUF files. In such cases, individual country summary statistics reported in the compendia may differ slightly from statistics used for international reporting.

INTERNATIONAL REPORT TABLES These report tables are publication-ready tables that support the international report. Typically, approximately 350–400 international report tables of varying length and complexity are generated. These tables can be used directly by the sponsoring agency or by countries as a means of quality control if they choose to conduct their own analyses and table production.

Preparation of the Technical Report

The technical report summarizes and describes all aspects of the study, including development of the frameworks for the background and context questionnaires, as well as for the cognitive domains; design and development of the survey instruments; development of the computer platform to manage, develop, and deliver the survey; translation and verification processes and procedures; survey operations and quality control procedures; sampling and weighting, data collection, and data processing; scaling, analysis, and preparation of data products; and reporting the results.

The technical report also includes any additional information related to the implementation of the psychometric and statistical methodologies at a level of detail that allows researchers to understand and replicate the analyses. To the extent possible, the report also addresses a range of common questions that both more and less advanced users of these complex databases might pose.

Tools and Training

A number of publicly available tools have been developed for use by countries, sponsoring organizations, and secondary researchers. For ILSAs sponsored by the OECD, two tools are available: One is a Data Explorer that allows users to navigate a secure, hosted database that includes all assessment cycles. The second is the IEA's International Database Analyzer (IDB Analyzer) that was adapted to work with OECD ILSAs. It enables local computer access and analysis of public-use and restricted-use databases.⁴²

Each tool addresses a slightly different set of needs and audiences. The Data Explorer is a web-based application that is usually hosted either by the sponsoring organization or by a government entity. It protects the underlying database while allowing relatively easy point-and-click navigation of the database to generate publication-ready summary cross-tabular tables and graphics.

The IDB Analyzer can be used in conjunction with restricted-use or public-use versions of the international database and uses microlevel data stored in a local computer (IEA, 2021). The IDB Analyzer satisfies the three main technical requirements that an analysis of these data must meet. They are (a) the use of sampling weights, (b) accommodating in variance estimation the complex multistage cluster sample design that was implemented, and (c) the use of multiple imputed proficiency estimates (i.e., the PVs).

Most commonly, the analyzer is used to generate data tables and graphs, as well as to execute either normal theory regressions or logistic regressions. The primary benefit of employing this tool is that it efficiently performs the required analytical computations, thus relieving the user of the programming burden. This process is especially critical in the computation of variances of estimates. In addition, the tool is accompanied by sufficient technical information and references to support both more- and less-advanced users in performing statistically correct analyses of the complex survey sample data.

To enhance data utility, sponsors and contractors organize and conduct workshops in various venues with different cost structures. The workshops focus on understanding the structure of the international database, the use of the IDB Analyzer and the Data Explorer, or other tools including standard statistical software. The participants in these workshops include staff from each participating country and interested representatives from the sponsoring agency. The sessions are arranged and coordinated by the contractor using staff members who are involved in or knowledgeable about the project design, the database preparation, and the appropriate analysis of the international data. The goals of the workshops are to help the representatives from each participating

country to understand the strengths and limitations of the data and to ensure that they are comfortable using the analytic tools that will be available and will report the results of the analyses with the relevant context and appropriate cautions.

In addition, the contractor or other organizations offer countries or interested researchers a fee-based option to organize and conduct regional workshops based on country-specified topics dealing with the technical aspects of using and analyzing a particular database. Additional workshops, presented under the auspices of the IEA–ETS Research Institute, the Institute for Education Sciences (U.S. Department of Education), and international donor organizations, are intended for interested researchers to expand the pool of knowledgeable users.⁴³ For example, the Institute for Education Sciences funded a multiyear set of training workshops on the use of the PIAAC data that resulted in dozens of research projects and publications. In addition, the IEA has constructed a website containing information about many ILSAs.⁴⁴

Data Security and Confidentiality

Data protection and confidentiality are essential to the conduct of all major ILSAs. Organizations wishing to conduct these types of assessments must comply with the national requirements of participating countries, along with international requirements associated with entities such as the European Union. Typically, all staff who work with these databases are required to sign a statement that they have read the information protection policy and that they understand and agree to abide by its provisions.

A growing number of organizations have adopted the International Organization for Standardization (ISO) 27000 series as their information security framework. ISO 27000 is an international series of standards that drives all aspects of the information security program and the way services are delivered. In the future, any organization wishing to develop or implement ILSAs will have to demonstrate that it is following applicable legal and regulatory obligations, with particular attention to the protection of personally identifiable information.

Typically, ILSA contractors have adopted stringent security policies, standards, and practices to maintain the confidentiality, integrity, and availability of data. These controls are continually reviewed to ensure they align with industry best practices. In addition, contracting organizations need to support secure transfer protocols, including file-based technologies to securely exchange data over the Internet. This enables them to effectively manage and protect data transmissions with secure file transfer protocol.

Data Quality

International large-scale assessments are complex surveys that require the development of a set of procedures for all major phases to help ensure that the sources of bias and variability in survey results are kept to a minimum so that reported results are relevant, comparable, and interpretable. Here, we distinguish two ways in which large-scale assessments address issues of data quality: the technical standards and guidelines that are developed and/or refined at the beginning of any assessment and the data

adjudication process that is conducted at the completion of the data collection and analysis process.

Technical Standards and Guidelines

At the beginning of each ILSA cycle, the ILSA team collaborates with the sponsoring organization and participating countries to develop and/or refine a set of technical standards and guidelines to ensure that all aspects of the survey satisfy accepted quality assurance guidelines. These standards and guidelines represent policies or best practices that must be adhered to in the development and conduct of the survey. Where compliance is not possible, countries may apply for derogations from the standards, as long as they are judged not to compromise the survey objectives. The overarching goal is to minimize total survey error (see the section “Population Sampling Operations”). The standards and guidelines cover a wide range of topics, including ethics, sampling, instrumentation, translation and adaptation, the use of hardware and software (if it is a technology-based assessment), data collection and training, data capture and processing, and data confidentiality and security.⁴⁵

Data Adjudication

Compliance with the technical standards and guidelines is an important component of ensuring the quality of the national data through the various stages of the workflow. At the conclusion of the workflow, yielding the databases described previously, data adjudication is implemented. Its objective is to render a judgment regarding the overall quality and fitness of the data of each country and to impose, if necessary, any limitations that should apply to the public dissemination and use of these data. In other words, the intent is to go beyond compliance with the standards and guidelines and determine whether the national data are of a sufficient quality to support their intended interpretations and uses. Adjudication typically involves three key areas: sampling and coverage of the target population, data collection and instrumentation, and quality of the translation/adaptation procedures.

For the OECD programs, the contractors develop an adjudication report that is reviewed and discussed at the final meeting of the Technical Advisory Group. If questions arise during the meeting, a sampling referee is available to offer advice as needed. The Technical Advisory Group then prepares a final set of recommendations to the sponsoring organization regarding the fitness of the data for publication. For IEA programs, the adjudication is carried out by a team comprising the different contractors and IEA staff.

IMPACT

Messick (1987) argued that large-scale assessments are a type of policy research and should be judged as such. In that light, if we regard policy utility as having the potential to inform, influence, or impact policy, then it is reasonable to ask to what extent—and

in what ways—that potential has been realized by ILSAs. As noted in the section “The Messick Framework, Extended,” the overall judgment of the potential for policy utility is dependent on evaluating the strength of an ILSA with respect to the design criteria of comparability, interpretability, and relevance—along with due regard to purpose, context, and the population(s) of concern. For the most part, users of ILSA results assume that the level of comparability is at least adequate for their purposes and that they (or their staff) have reached appropriate and defensible interpretations of the data. Ultimately, the judgment of relevance provides at least part of the motivation to cite the data in discussions and policy debates. However, the contexts in which ILSA data are employed and their actual impact on education policies and practices are the result of complex dynamics that vary widely over time and space. Thus, the evaluation of impact requires considering a broad range of cases from which a few generalizations can be gleaned.

Background

To best address the question of impact, it is helpful to both distinguish different types of impact and then systematically evaluate the various kinds of evidence that are available. In this regard, D. Rutkowski et al. (2020) presented a framework for evaluating ILSA influence on policymaking. The framework comprises a causal model for tracing the impact of ILSAs on a particular policy and a logic model for evaluating the evidence for such impact.

Drawing on D. Rutkowski et al. (2020), we define impact as occurring when it is possible to connect (in some way) ILSA data, technical reports, or secondary analyses to changes in policy and/or practice in a participating country or even shifts in the content of conversations about policies in education, workforce development, or labor market regulation at the national or cross-national levels. However, as Wagemaker (2014) pointed out in discussing school-based ILSAs, it can be problematic to judge policy impact because

major policy initiatives or reforms are more likely to result from a wide variety of inputs and influences rather than from a single piece of research. Research is also more likely to provide a heuristic for policy intervention or development rather than being directly linked, in a simple linear fashion, to a particular policy intervention. (p. 12)

Two further challenges are (a) a frequent time lag between ILSA reports and policy decisions and a further lag between policy initiation and its effect (if any) on measured achievement or skills and (b) variations in policy implementation across the target units within a country that can reduce estimates of policy impact (Braun et al., 2006; Burdett & O'Donnell, 2016). Thus, correctly judging impact involves triangulating different kinds of evidence, including statements of key actors, stakeholders, and others as to the role of ILSAs in policy decisions, formulations, and implementation (Fischman et al., 2019). Accordingly, we discuss two types of impact: direct and indirect.

Direct Impact

Direct impact occurs when results released to the public (appear to) have a relatively clear, and even immediate, influence on policy makers' discussions, priorities, and actions. Some actions may address governance and funding; others relate to curriculum, teacher training, pedagogy, and assessment; and yet others involve achieving greater harmonization across different units within the education system.

The literature on school-based ILSAs is replete with examples of how ILSA results precipitated intense discussions that led to substantial policy changes. Perhaps the most celebrated example is the impact of PISA 2000 results on Germany (Ertl, 2006). The relatively poor performance of German students (following their poor performance on TIMSS 1995) resulted in unprecedented cooperation among the federal government and all the German states, leading to changes in curriculum and assessment. The need for such changes was supported by results from PIRLS 2001, which indicated that children from immigrant families on average performed substantially more poorly than their peers. In response, the national research agenda shifted toward more empirically based studies with direct relevance to practice (Schwippert, 2007).

However, Pons (2017) suggested that this type of analysis may be too simple. He argued that before reaching a judgment on the "PISA effect" (or its equivalent), it is necessary to consider multiple factors, including the prior policy context and education-related political debates. On this point, he asserts, "PISA uses in the policy process greatly depend on the dynamics of the domestic policy debate and on the preexistence of specific structuring controversies that PISA results illuminate in a new way" (Pons, 2017, p. 139). This is consistent with the perspective offered by Ritzen (2013) and Braun (2013). Indeed, there are many types of "policy borrowing," ranging from fairly direct adoption to adapting certain policy features as part of a broader policy formulation (Burdett & O'Donnell, 2016).

Heyneman and Lee (2014) cited additional examples of direct impact in countries as diverse as Denmark, Macedonia, Kuwait, and Japan. They also provided an extensive list of direct impacts on policies in other countries, as well as on subject-specific policies and practices. Less noticed, but perhaps equally important, ILSA results have also been used to provide support for existing policies. Heyneman and Lee (2014) cited the example of PIRLS 2001 in England. The relatively strong performance of English students was used as evidence of the effectiveness of the National Literacy Project that was introduced in 1996. Similarly, the performance of students in Australia and New Zealand on PISA has been taken as support for existing educational policies. Breakspear (2012) called out countries such as Singapore and Spain that use PISA to complement (and to validate) the results of their national assessments. Canada participates in most ILSAs, often with oversampling, and makes systematic use of the results to guide policy at the national and provincial levels (Volante, 2013).

Singapore represents an interesting case study (Ng et al., 2020). Singapore has participated in IEA studies since the 1990s and employs ILSA results to inform curriculum and pedagogy, as well as to evaluate the efficacy of various reforms. The authors stress,

however, that these results are only one of many sources of evidence that policy makers and education leaders consider in deciding on policy strategies and education initiatives. They also note that to discourage score corruption, the Ministry of Education uses the scores only to inform system-level decisions and not for any form of accountability.

Baird et al. (2016) conducted an intensive study of the impact of PISA on six countries.⁴⁶ Based on both a comprehensive document review and multiple interviews, they concluded that in five of the countries (excluding England), some of the policy changes could reasonably be attributed to the impact of PISA results. For example, the two federal systems (Canada and Switzerland) initiated a process of greater harmonization of achievement standards and assessments across provinces and cantons, respectively. In the case of France, PISA prompted substantial curricular changes, as well as initiation of a sample-based national assessment modeled on the ILSA design.

With regard to PIAAC, there is little evidence of direct impact. On the one hand, this may be because different ministries focus on one or another aspect of the results and then evaluate them in light of other relevant information, as well as political considerations. On the other hand, the assessment of adults' cognitive skills, as well as their distribution at different levels of educational attainment, has proven revelatory.

Indirect Impact

Indirect impact occurs in a variety of ways. First, ILSA findings may be used to justify, legitimate, or build public support for policy prescriptions that have already been formulated and even initiated (Fischman et al., 2019). Second, the results of one or more ILSAs over multiple cycles can change the nature of the discourse regarding important aspects of education, skills, and skill development. In addition to the achievement results themselves, subsequent to the release of the PUFs, analysts delve into the ILSA databases to investigate patterns of relationships and how they differ across countries and over time.

The comparative findings can influence policy discussions regarding the need to remedy apparent deficits or inequities. In other cases, researchers attempt to draw inferences regarding the efficacy (or lack thereof) of particular policies or constellations of policies with regard to certain outcomes. Yet others seek to draw attention to possible unintended consequences of employing ILSA data in policy discussions. Such secondary analyses, whatever their nature, appear years after the initial release of the data. Nonetheless, the findings and interpretations may also influence the discourse related to particular policies and, in some instances, even provide the impetus for changes in policy or practice.

As noted earlier, ILSAs have experienced increased participation by medium- and low-income countries. Undeniably, there is some element of coercion by funders. However, in addition to providing critical policy insights and guidance, there are many other salutary benefits. These include constructive changes in regulatory policies (e.g., curriculum content, performance standards) and in so-called behavioral policies (e.g., pedagogy, educator professional development; Lockheed, 2013; Wagemaker, 2014).

A particularly important benefit associated with participation in school-based or adult assessments is the contribution they make to capacity building. By participating in an ILSA, country representatives and technical experts receive extensive training and support in all aspects of managing and conducting large-scale assessments (Meinck et al., 2020). These experiences and the expertise gained can then be used not only in future ILSAs, but also in improving (or initiating) national assessments that are directly targeted at informing education policies of greatest priority. Lockheed (2013) cited numerous self-reports of such impact.

Since 2003, participation in the ILSAs that focus on adults has grown in tandem with a marked shift in focus among policy makers who have moved from viewing educational attainment as an appropriate proxy for skills to recognizing the added value of comparable and valid/relevant measures of skills of interest.⁴⁷ One reason is the accumulating evidence that there are substantial economic returns to skills beyond those accounted for by educational attainment or years of education (Fogg et al., 2018; Hanushek et al., 2015; Kirsch et al., 2007). Such findings not only enhance policy interest in tracking skills, but also highlight the societal perils in tolerating extreme inequalities in skill distributions. Consequently, in many countries, finding ways to accelerate skill growth while reducing inequities have become policy imperatives—fueling greater interest in studying the policy initiatives of “high-flyers” (i.e., those countries at the top of the league tables or those who have experienced large gains over two or more cycles).

Indeed, greater attention in the media and the exponential increase in Internet searches related to ILSAs are testament to their growing role around the world in discussions of education policies and practices. As Ritzen (2013) pointed out, the comparative outcomes derived from ILSAs make these assessments powerful instruments of transparency. Their findings can jumpstart (or accelerate) national conversations regarding quality and equity in education or in labor market dynamics. Over time, such conversations, particularly at the policy level, can lead to shifts in perspectives, new understandings, and the realization that far-reaching system changes are not only possible but also much needed (Conaway, 2020).

Secondary Analyses: Exemplars

There has been an increase in the number and range of researchers using ILSA results for studies addressing important policy issues, with a concomitant increase in the utilization of the IEA's IDB Analyzer. Pons (2017) provided a useful review and analysis of articles employing PISA data. With respect to adult assessments, Maehler et al. (2020) compiled a bibliography of published papers between 2008 and 2019 covering a range of topics including research results associated with PIAAC assessment. Since 2013 when the PIAAC data were first released, the number of publications has increased annually. The current bibliography contains more than 600 publications and 21 technical reports. More generally, the IEA gateway (<https://ilsa-gateway.org/studies/papers>) contains a comprehensive list of articles employing both IEA- and OECD-sponsored ILSAs. For a more recent integrative review, see Hernández-Torrano and Courtney (2021).

Among the issues considered in these analyses are skill gaps within and among participating countries. Of particular import are studies comparing countries with respect to the gradient of performance against a measure of socioeconomic status.⁴⁸ Although a gradient exists in every country among school-age children (i.e., mean performance being positively correlated with a measure of socioeconomic status), there is a wide range of gradient values, some strikingly large (OECD, 2019, chap. 2). As it happens, the gradient in the United States is typically the largest among OECD countries.

There are similar findings for adult populations. For example, Goodman et al. (2015) examined the skills distributions and inequality among millennials in the United States with respect to their overall performance, as well as in comparison to other countries participating in PIAAC. For the United States, the authors considered differences in skill distributions by factors such as educational attainment and race/ethnicity. With regard to gender differences, Braun (2018), employing PIAAC data, demonstrated that in every participating OECD country the earnings of women employed full-time trailed those of men employed full-time, even after adjusting for age, family background, cognitive skills, educational attainment, and occupational category. Again, the degree of disadvantage varied widely across countries. In an empirical study using PIAAC data, Vera-Toscano et al. (2017) demonstrated that, over and above educational attainment, measured cognitive skills have incremental predictive validity for social outcomes such as participation in volunteering activities.

Strietholt and Scherer (2018) argued that combining different ILSAs or linking ILSAs to other data sets increases the range of research questions that can be addressed. They cite, for example, the study by Martin et al. (2013) that took advantage of the conjunction of PIRLS and TIMSS in 2011. Through careful planning, 34 countries and 3 benchmarking countries administered PIRLS and TIMSS to the same fourth-grade students. Thus, each student generated data on skills in reading, mathematics, and science and responded to background, attitudinal, behavioral, and school context questions. Among other things, the authors found substantial heterogeneity across countries in the amount of variation in the cognitive outcomes accounted for by various background variables. By contrast, within countries they found typically small differences in the amount of variance explained across cognitive domains.

Strietholt and Rosén (2016) offered an example of combining different assessments over time. Specifically, they linked the Reading Comprehension Study 1971, the Reading Literacy Study 2001/1991, and PIRLS 2001, 2006, and 2011 to study reading trends over a span of 40 years.⁴⁹ Hanushek et al.'s (2015) study cited previously involved combining data from PIAAC with income data obtained from various national and international sources. In summary, the results of cross-national, comparative assessments, as well as the studies that employ those results, can provoke useful discussions among researchers and stakeholders that likely would not otherwise take place.

Technical Issues in the Study of ILSA Impact

As this section and the references therein have amply documented, ILSAs have had considerable indirect and even direct impact on education and skill development policies around the globe. However, the process of policy formulation, adoption, and implementation has a strong political component that can dominate the influence of ILSA data and related secondary analyses. Indeed, as already noted, impact depends on the readiness of the relevant national actors to take account of the information provided, to decide on which action(s) to take, to articulate policies and plans, and to commit both the necessary funding and the political capital to implement these policies. ILSA information is just one source of evidence (or pressure) among many that can motivate or drive educational change.

Notwithstanding these political realities, technical considerations should (and sometimes do) urge caution on ILSA interpretations and impact. In the early 2000s, critics of ILSAs typically focused on technical deficiencies related to sampling and translation/adaptation that, in their view, substantially reduced ILSA utility (Bracey, 2008; Goldstein, 2004). Later, focusing on PISA 2012, L. Rutkowski and Rutkowski (2016) undertook an analysis of these technical issues with regard to their implications for defensible interpretations of the results and the use of those results in policy discussions. Their principal recommendation was that sponsors should be fully transparent regarding the bias and uncertainty (variance) associated with published results and their implications for the interpretation of such results. Despite the significant improvements in PISA technical quality achieved since then, as well as those realized with other ILSAs, the cautions and recommendations offered by the authors should be borne in mind by both sponsors and users of all ILSA results.

For example, the assumption of cross-national comparability of domain scores and responses to the background questionnaire is only an approximation (see the section “Measurement Invariance”). Consequently, conclusions should be tempered by an acknowledgment that for a particular analysis the assumption may not hold for some countries to an extent that renders the conclusions suspect, at least for those countries. Such cautions may limit the impact or even the influence of the study. On this point, it is beneficial (if not essential) that participating countries have the “in-house” expertise to analyze their data, to explain the findings to various stakeholders, and to contribute constructively to policy conversations. Participation in ILSAs has stimulated many countries to invest in developing such expertise, along with the infrastructure required to support high-level methodological research.

More broadly, success in policy borrowing is difficult to achieve. Education systems are complex: They comprise many component systems with complicated and shifting dynamics. Identifying a particular policy or practice as a key driver of academic achievement may neglect concomitant factors or conditions that are essential to its effectiveness. Even if the identification is approximately correct, adapting the policy or practice to a setting that differs in many relevant ways (e.g., power dynamics, traditions and culture of schools, resources) is

challenging. In the course of technical adaptation and responding to political pressures, the essence of the innovation may be “lost in translation” (Burdett & O’Donnell, 2016).

Furthermore, the information provided by an ILSA is typically more narrowly focused than the general questions posed by policy makers (D. Rutkowski & Delandshere, 2016). For example, policy makers would like an answer to the question of whether observed differences in achievement among countries can be attributed to differences in education policies. However, as pointed out by Braun and Singer (2019),

Although it is the kind of question policy-makers often view as most relevant, most methodologists agree that ILSAs are ill-equipped to provide unambiguous answers. Cross-sectional data based on samples without random assignment of subjects into experimental and control groups do not lend themselves to valid causal inferences. Attempting to infer causation from correlation, even when correlations are high, can lead to false, incomplete, or misleading conclusions. (pp. 79–80)

Because the successful implementation of educational policies across so many different contexts depends on a broad range of relevant factors, both educational and otherwise, it is exceedingly difficult to disentangle the influence of these factors with cross-sectional, observational data. Table 20.2 displays six possible uses of ILSA results with judgments on the general suitability of ILSA data for such uses. Evidently, suitability declines as the uses shift from description to inference.

Table 20.2 Purposes of School-Based International Assessments

Item	Purpose	Capacity of ILSAs to Achieve This Purpose
1	To disturb complacency about a nation’s education system and <i>spur education reforms</i> .	Outstanding
2	To <i>describe</i> and <i>compare</i> student achievement and contextual factors (e.g., policies, student characteristics) across nations.	Excellent
3	To <i>track changes over time</i> in student achievement, contextual factors, and their mutual relationships, within and across nations.	Excellent
4	To <i>create de facto international benchmarking</i> , by identifying top-performing nations and countries, or those making unusually large gains, and learning from their practices.	With caveats
5	To <i>evaluate</i> the effectiveness of curricula, instructional strategies, and education policies.	With extreme caution
6	To <i>explore causal relationships</i> between contextual factors (demographic, social, economic, and educational variables) and student achievement.	Dangerously difficult

Note. ILSA = international large-scale assessment.

D. Rutkowski and Delandshere (2016) argued that causal claims based on ILSA data should be critically evaluated through the lens of a validity framework comprising four facets: construct validity, internal validity, external validity, and statistical conclusion validity. The findings of such an evaluation can be quite sobering, as the examples cited by the authors attest. In general, causal claims can be buttressed both by an *a priori* theory of action and by an argument of the comprehensiveness of the observed variables included in the analysis. Nonetheless, such claims must remain tentative until other supporting evidence comes to light (Singer et al., 2018).

That said, even though ILSAs rarely permit causal inferences, their findings can reveal striking patterns, raising issues that can be carefully examined through further empirical studies. This caution is particularly relevant in moving from tentative causal descriptions to credible causal explanations, with the latter being more relevant to policy decisions and actions (D. Rutkowski & Delandshere, 2016). In this regard, small-scale experimental studies that demonstrate the importance of certain interventions or strategies can then be more broadly validated within and across national samples.

Some Extensions to Discerning Impact

It is sometimes possible to take advantage of auxiliary information to circumvent the limitations of cross-sectional studies. One strategy employs an instrumental variable to obtain an estimate of a causal effect. Unfortunately, it is usually difficult to find a suitable instrument, especially in an international context. Pokropek (2016b) presented one example using data from Poland. Another example was offered by Bedard and Dhuey (2006). They employed school entry cutoff dates as an instrument to estimate the impact of delayed entry into schooling (see also Marchionni & Vazquez, 2019). In other settings, regression discontinuity designs may also prove useful (Robinson, 2014).

A different approach is to conduct true longitudinal studies in which individuals participate in related assessments on two or more occasions. Differences in outcomes can be linked to differences in treatments (or other factors) to yield causal estimates with some degree of credibility. Carrying out such studies always poses substantial technical, logistical, and sometimes ethical challenges, especially in an international context. One strategy is to construct a pseudo-longitudinal study by concatenating different assessments suitably spaced in time. Kaplan and McCarty (2013), employing data for PISA and Teaching and Learning International Survey (TALIS) from Iceland, examined a number of methods for creating a synthetic data file. At the national level, there are a few instances of such studies. In Denmark, for example, a subsample of the PISA 2000 sample was interviewed and tested as part of PIAAC 2012. By linking the assessment data to data from administrative registers, the author was able to account for some of the variation in individual rankings between assessments in terms of both fixed and varying characteristics of the individuals (Rosdahl, 2014). However, the possibility of construct shift suggests caution in interpretation.

An alternative is to combine data from successive cycles of the same assessment (e.g., PIRLS 2016 with PIRLS 2021) and relate changes in the assumed causal agent with observed changes in achievement. This feature has been exploited by Gustafsson (2013), Gustafsson and Nilsen (2016,) and Hooper (2017). The basic idea, attributable to Gustafsson, is to employ the difference-in-differences methodology at the national or subnational level. Needless to say, the validity of the causal claims depends on building the counterfactual case that changes in achievement can only be due to the changes in the hypothesized causal agent. In the differences-in-differences methodology (Cunningham, 2021), other plausible causes are discounted by means of the parallel trends assumption, where trends between similar populations (e.g., countries) are expected to be parallel if an effective causal agent is not present. Parallel trends between similar countries (e.g., Sweden, Norway) occur frequently in economic data where higher level economic trends (e.g., global economy) have a strong influence on lower level economies (e.g., national economies). Parallel trends are less common in international assessment data, where cycle-to-cycle fluctuations in achievement results do not tend to follow a similar pattern across countries, making it more difficult to assume that deviations from parallelism are due solely to the presumed cause. Given the demand for credible policy evaluation, these and other methods are the subject of continuing research.

The study of correlates of student achievement is of long-standing interest. Because TIMSS and PIRLS sample students by class, collecting data on students and their teachers, they provide an opportunity to link teacher characteristics and teacher practices to student achievement at an international level. However, because student achievement is cumulative, its association with teacher characteristics in a particular year may be quite weak. Comparisons among countries with regard to both teacher practices and student achievement can be suggestive of directions for further investigation. O'Dwyer and Paolucci (2019) discussed what has been learned, along with a careful analysis of the obstacles to making inferences about causal relationships.

The challenge of addressing questions regarding the efficacy of reform initiatives with ILSA data has engendered considerable ingenuity among methodologists, leading to new analytic strategies. Similarly, as the section “Transitioning to Digitally-Based Assessments” amply demonstrates, ILSA teams have harnessed emerging technologies to construct innovative platforms to carry out the processes that undergird the ILSA programs. These innovations have enabled the ILSA teams to meet the ever-increasing demands of sponsors and countries while meeting constraints of time and cost. Both developments exemplify the virtuous spiral (Figure 20.2). In the section “History of International Assessments” we noted that the ETS proposal to reinvent NAEP (Messick et al., 1983) represented a creative response to policy makers’ demands for NAEP to provide data that were both more relevant and more interpretable. Certainly, as the global landscape continues to evolve and as ILSAs yield more useful insights, new demands will arise necessitating further innovation. Accordingly, the following two sections discuss some of the technical and political challenges that ILSA teams will likely face in the coming years.

TECHNICAL CHALLENGES

Notwithstanding the remarkable growth in ILSA participation since the 1990s, as well as the bright prospects for ILSAs with a future based on digital platforms, program sponsors and contractors will be confronting a number of technical challenges to policy utility. Many of these challenges arise from the need to maintain (or even enhance) relevance as the context of use and the intended applications continue to evolve. With the corresponding changes in the assessment frameworks, instrument development, and mode of administration comes the need for additional evidence to support validity claims.

Indeed, an underlying theme of this chapter has been the assurance of validity of interpretation and use within a policy setting. A judgment of validity must be made with due regard to the context of use and the particular purpose(s) for which the results will be employed (Kane, 2013, 2016). Much of the relevant evidence is contained in the technical reports that accompany each ILSA and has been discussed in various sections of this chapter. The section “Transitioning to Digitally-Based Assessments,” for example, describes the key phases of the workflow processes, including the specification of the frameworks that guide the development of the instruments and the accompanying quality control procedures that are applied during assessment development, translation/adaptation, and implementation. These methodologies, along with the data generated through monitoring their implementation, constitute evidence for “procedural validity,” forming a foundation for a validity argument.

Similarly, the section “Scaling, Population Modeling, and Proficiency Estimation” provides an overview of the statistical and psychometric models and procedures that transform the raw data into the published results. The mathematical and verbal representations of these models and procedures make explicit the strategies employed to take account of ILSA design features (e.g., planned missingness, group-level reporting) in generating the results. Thus, these strategies can be—and have been—subjected to critical review and ongoing refinement to support the validity claims.

The section “Policy and Political Challenges” addresses some of the political challenges ILSA sponsors and contractors face with the greater salience of ILSA results in national and global educational policy debates. Much of that discussion can be framed in terms of the consequential validity of ILSAs, which is essential to the role of ILSAs as policy research and can sometimes become politicized and highly charged. In contrast, this section focuses on issues having more to do with construct validity (Messick, 1989).

We now review some of the technical challenges confronting ILSAs, using the extended Messick framework to organize the discussion. At the same time, it is important to recognize that challenges represent not only problems, but also opportunities. By appropriately meeting these challenges, ILSAs can continue to traverse the virtuous spiral (Figure 20.2), developing in ways that enhance their utility and impact. We begin by considering direct threats to comparability and interpretability.

Threats to Comparability and Interpretability

One threat, as noted previously, is related to the likelihood that ILSAs will have to maintain dual systems (paper-and-pencil administration and an ever-evolving digitally-based assessment) for some time—both because of insufficient technical capacity in many newly participating countries and because segments of some target populations (e.g., older adults) may not be capable of responding with tablets or other digital devices. To this point, through statistical adjustments based on estimated mode effects, ILSAs have maintained a single reporting scale for results obtained through either paper-and-pencil administration or digitally-based assessments. The construction of a single scale implies that the two sets of scores are linked to substantially the same construct and support similar interpretations.

Over time, however, ILSAs employing a digital platform will continue to innovate by extending the number and scope of construct facets to reflect the broadening of legacy constructs or the introduction of new constructs that capture the ways in which students and adults interact with new technologies. To some extent, these changes will involve new item types and new response formats that leverage the affordances of the technology platform. This will lead to a growing divergence between the paper-and-pencil administration and digitally-based assessment instruments, thereby making the assumption of the comparability of the scores obtained in the two modes progressively less tenable. Consequently, interpreting the scores (and score differences between countries) without regard to mode will become very problematic—especially if the proportions of respondents in the two modes differ substantially across countries.

We can expect that, eventually, separate scales will be required, with the attendant complications for reporting and interpretation. Moreover, the additional cost of maintaining two different platforms is not trivial. For example, one continuing challenge concerns transitioning paper-and-pencil administration trend items to digitally-based assessments.⁵⁰ This requires careful adaptation to keep the information provided by the item in the two formats comparable (Lennon & Kirsch, 2025; von Davier, Khorramdel, et al., 2019).⁵¹ Finally, within the digitally-based assessment orbit, the proliferation of different delivery devices over time will also pose a challenge to comparability both within and across countries.

A second threat concerns the shift to adaptive multistage testing. This move necessitates significant increases in the sizes of the item pools, particularly with the enhancement of construct representation. In conjunction with the growth in the number of participating countries, these increases place greater demands on the capacity of ILSA teams, not only to develop sufficient numbers of items with desired characteristics, but also to carry out the operations of translation and adaptation with the high degree of quality essential for comparability. Furthermore, greater numbers of items to be calibrated necessitate increases in the sample sizes required and/or the number of items administered to each respondent during field trials.

As documented in the section “Transitioning to Digitally-Based Assessments,” ongoing innovations in tools and processes are required to meet these demands, all the while respecting severe constraints on cost and time. Furthermore, adaptive multistage testing substantially increases the complexity of assessment design and, consequently, heightens the need to develop new quality control procedures to ensure that the overall design requirements are properly satisfied. In this regard, automated test assembly software that can be implemented for adaptive testing should prove very useful (Luo, 2020).

Third, the introduction of innovative item types may not only drive a divergence between modes of administration, but also introduce nonnegligible construct-irrelevant variance due to differential familiarity with response formats within and especially across countries.⁵² Whether a nonnegligible fraction of the observed variance is deemed construct relevant or construct irrelevant bears on how to address the issue of measurement invariance (see the section “Measurement Invariance”).

Finally, an ongoing concern will be ensuring an adequate level of measurement precision and comparability of the scores in the cognitive domains in the face of increasing heterogeneity among participating countries with respect to both cognitive skills and relevant background knowledge.⁵³ L. Rutkowski and Rutkowski (2018) discussed a relevant paper-and-pencil administration strategy, termed *design based*, that involves creating booklets with different average levels of difficulty and adjusting the proportions of relatively “easy” and relatively “hard” booklets administered in a country based on an a priori estimate of the proficiency distribution in that country. Indeed, such a group-level, adaptive strategy has been applied on a limited basis in PISA 2009 and PIRLS 2021.

Questions of comparability also pertain to the scales developed from responses to the background questionnaire. The challenges are greater in this setting than with the cognitive scales inasmuch as the number of items contributing to each background questionnaire-derived scale is very small. Consequently, measurement error is substantially greater. Furthermore, in some countries, scores on the reading/literacy assessment indicate that many (if not most) of the students may have some difficulty understanding the questions and responding appropriately.

L. Rutkowski and Rutkowski (2018) also addressed issues of comparability of the scales derived from the background questionnaire. They described a strategy, termed *model based*, that involves somewhat relaxing the measurement invariance requirements (to partial invariance) and encouraging countries to take advantage of the so-called national option to augment the core items for select background questionnaire constructs with items of specific interest to the country (or to a group of countries). The intent is to enhance interpretability and utility. This strategy was demonstrated using the WEALTH scale of PISA 2012. The exercise illustrated the challenge in achieving cross-national comparability even with a group of similar countries.⁵⁴ At the same time, there is a trade-off with comparability across all participating countries. For further empirical analysis, see, for example, He et al. (2019).

Van de Vijver (2018) considered the variation in response styles across countries and cultures particularly problematic with the Likert scale response formats used in many background questionnaire items. In this regard, Kyllonen and Bertling (2013) described some of the problems that arise with items that employ typical rating scale response formats and investigate some alternative response formats. Some formats tend to yield results that exhibit the aptitude–achievement paradox.⁵⁵ This is an area of ongoing research (von Davier et al., 2018). A comprehensive treatment of response format issues in an international context can be found in Kuger et al. (2016). A more general discussion of the validity issues related to items requiring self-report can be found in Karabenick et al. (2007).

As noted in the discussion of translation and adaptation procedures (see the section “Translation/Adaptation/Verification of the Instruments”), numerous efforts exist to enhance cross-cultural equivalence in the measurement of the target constructs of the background questionnaire in the face of greater heterogeneity among participating countries. Nonetheless, credible concerns remain (L. Rutkowski & Rutkowski, 2019). In addition to bias, another consequence is that some countries experience higher levels of measurement error in the scales derived from the background questionnaire that, in turn, attenuate the estimates of the relationships between cognitive skills and the constructs underlying those scales. These issues, singly and in combination, can substantially reduce the amount of useful information generated for some countries.

Notably, cross-national differences in measurement error in background questionnaire scales can result in spurious differences in correlations between such scales and cognitive domain scores. An important question, then, is to what extent observed heterogeneity across countries is due to such statistical artifacts. Gurkan (2021) developed an approach to correcting the bias in correlational estimates that takes account of both the multilevel structure of ILSA data and differential measurement error. The method further posits that the tested population comprises two latent classes with different levels of domain proficiency. When applied to PISA 2015 data to examine the relationships between mathematics proficiency and mathematics self-efficacy, the method resulted in substantially reduced correlational heterogeneity across countries, especially for lower performing countries.

A fifth threat, primarily to school-based surveys, is due to the variation among countries in the ways they identify students with different types and degrees of disability, their rules for exclusion from the survey, and the nature of the accommodations employed during survey administration (see the section “Accommodations for Testing”). On the one hand, some of the variation arises from different national customs and regulations and is difficult to mitigate. Certain subpopulations, such as recent immigrants, may also be treated differently across countries, again resulting in some lessening of comparability. On the other hand, some sources of between-country variation may be susceptible to harmonization, that is, coordinating definitions of factors such as education levels and socioeconomic status. Efforts to that effect in the social science literature

are ongoing, with the goal of increasing comparability of patterns of relationships across countries.

A sixth threat arises from the proliferation of mandated, official school surveys, as well as school-based research studies. This has led to “survey congestion,” resulting in lower participation rates. Although replacement schools can be located and convinced to participate, the overall trend is worrisome because, over time, there can be greater divergence in sample quality among countries. In this regard, Durrant and Schnepf (2017) linked PISA sample data for England to two large-scale administrative databases with information on students’ socioeconomic backgrounds and performance on national public examinations. They were able to characterize both the schools and the students most likely not to respond and, on that basis, made suggestions on how best to select substitute schools. This approach shows some promise if it could be implemented in countries where such linkages are feasible.

In the case of household surveys, such as PIAAC, reduced participation rates may be due to both survey congestion and a general decrease in trust in central authorities. Here again, increasing divergence in sample quality threatens country-level comparability. Lowered participation rates may necessitate offering some incentives on a broader scale.⁵⁶ The COVID-19 pandemic and its aftermath may well exacerbate these challenges. It is possible that in the future some forms of remote assessment will offer a partial solution.

Even among those who agree to participate, there are worrisome trends in motivation. Here the concern is greater with school-based surveys than with household surveys. In low- or no-stakes situations, many students do not engage fully with the instrument—or only do so for part of the assessment. The consequence is less accurate measurement at the population and subpopulation levels, as well as reduced comparability, because diminished motivation will vary in prevalence across subpopulations within and across countries. For an estimate of the impact of differential motivation on NAEP results, see Braun et al. (2011). All these threats to comparability serve to reduce the policy utility of international large-scale assessments. Recent approaches to analyzing item-level data for engagement may prove useful in this context (Ulitzsch et al., 2019).

In the future, log file data should prove to be of use. For example, timing data are already helping in the effort to classify items at the respondent level as either omitted or not reached. Researchers have begun the development of indicators based on the raw log file data that can be used to signal possibly aberrant response processes or patterns suggesting lack of effort (e.g., Goldhammer et al., 2016; Pokropek, 2016a; Pools & Monseur, 2021; Wise, 2020). One strategy would be to develop norms for an indicator (possibly at a national level) and to identify individuals with indicator values that fall in the extremes of the distribution. If that is the case, say, for two or more indicators, then that individual would be removed before scaling and analysis. Of course, empirical work would be necessary to validate a decision rule (Soland et al., 2021). In any case, such indicators could be used to monitor system performance and enhance sample quality.

Another possibility is to employ records of participant response processes to improve item design and to contribute evidence of validity (Ercikan & Pellegrino, 2017; Zumbo & Hubley, 2017). Although these seem to be promising directions, many technical, practical, and even ethical questions remain to be explored (Provasnik, 2021). Technical advances and considerable infrastructure development will be required to make these and other possible applications of log file data practicable. Ethical issues should be considered by broad-based groups of stakeholders, with sufficient international representation, although the different perspectives may well make reaching consensus challenging in its own right.

Finally, comparability and interpretability of ILSA results also depend on whether respondents differentially interpret the tasks that are presented and how those interpretations shape their response behaviors. Assessment practitioners are coming to understand the importance of adopting a sociocultural perspective in the design of assessments, as well as in the interpretation of the data generated by those assessments (Mislevy, 2018). In the ILSA context, such considerations are germane to both the cognitive instruments and the background questionnaire. One implication is that the committees that develop these instruments should have broader participation from the different subpopulations in the various countries to enhance the meaningfulness of the questions to all respondents, thereby reducing construct-irrelevant variance.

Although the logistics of carrying out such a program in an international context are formidable, ILSAs sponsors and contractors do make an effort to include representatives from different countries as subject matter experts, to attend test development workshops, and to contribute items for field testing. Evidently, more needs to be done in this regard.

Relatedly, some critics have suggested that ILSA teams should be open to employing different methods for generating validity evidence. For example, Pepper (2020) pointed out that the validity argument would benefit from conducting cognitive laboratories with both the cognitive instruments and the background questionnaire items.⁵⁷ In an international setting it would be impractical to do so in every participating country. Consequently, some experimental design would have to be employed to generate evidence reasonably representative of the ensemble of participating countries. Classical convergent and divergent validity studies could also be conducted in selected settings. In any case, more thought should be devoted to these and other approaches to obtaining additional evidence. Similar considerations apply to studying the impact on the performance of certain subpopulations with the introduction of new digital devices.

Threats to Relevance

ILSAs also face challenges with regard to relevance. For the most part, the principal strategy to enhancing relevance has been to increase the range of constructs that are measured through the cognitive instruments and the background questionnaire. However, there are technical limitations to pursuing this strategy. Moreover, each cycle brings novel issues to the fore.

One aspect of relevance is the degree to which ILSA assessments reflect the increasing use of different technologies by students and adults. For example, there is ongoing interest in incorporating various 21st-century skills in school curricula (Binkley et al., 2012; Darling-Hammond, 2012). The assessment of such skills often requires complex stimulus materials and student-constructed responses that currently require human raters for evaluation. Including such (or similar) assessments in ILSAs poses a number of challenges, as has already been mentioned. Nonetheless, to maintain relevance, expanding the range of assessment formats remains a high priority for ILSA teams.

Fortunately, the advent of digitally-based assessments makes possible new item types with different response formats to enhance construct representation. As an example, simulation tasks are becoming increasingly popular in school-based ILSAs because they allow for a more authentic assessment in the domains of science inquiry, mathematics, collaborative problem-solving, and even financial literacy (PISA). In the case of science, simulations can provide students with an opportunity to design an experiment (by selecting variables and associated values) and then run the experiment to generate data in order to respond to various questions. Because such tasks can consume considerable time, their appearance in an ILSA may be more common in specialized auxiliary studies.

With the introduction of adaptive multistage testing, it is optimal to have decisions on the next module to be administered based on the maximum amount of information. Currently, information is only available from selected response items and a limited number of other item types that can be automatically scored. Going forward, new item types should be developed and evaluated in conjunction with the corresponding scoring algorithms. This shift requires item designers to collaborate with computer scientists and other specialists to determine the feasibility of developing accurate and efficient scoring algorithms. In some cases, original item designs will have to be modified with a commitment to maintaining construct relevance. Evidently, implementing automated scoring in an international context with multiple languages and scoring guides is an order of magnitude more complicated than in national assessments. Nonetheless, the implementation of adaptive testing in response to demands for greater accuracy along the full score scale, together with the necessary concomitant technical developments, is yet another instance of the virtuous spiral (Figure 20.2) that captures the ongoing dynamic between stakeholders on the one hand and ILSA sponsors and teams on the other.

On another front, ILSAs must be responsive to changes in the school, work, or other environments where the easy availability of online tools (e.g., for checking spelling, grammatical construction, argumentation, information search and display) has the potential to modify the focal construct. Maintaining both relevance and scale comparability across administrations may be increasingly at odds in both student and adult surveys. As Mazzeo and von Davier (2014) pointed out,

The question is whether surface characteristics that change quickly due to technological advances will lead to changes in the requirements of underlying skills and knowledge. If students increasingly use technology in everyday activities, and if

these technologies become easier to apply to everyday problems over time, then the traditional concept of linking assessments over time by means of tasks that stay the same and look the same becomes unsuitable. (p. 255)

This poses a challenge not only to instrument developers and psychometricians, but also to those charged with renovating the digital platform. Investments in continuous innovation in platforms, processes, and procedures will consume substantial resources. More fundamentally, technology-driven changes in the context of measurement can force shifts in how the construct is defined and operationalized. As a result, what may have once been viewed as a source of construct-irrelevant variance may come to be accepted as contributing construct-relevant variance. An example is provided by the construct problem-solving in technology-rich environments, introduced in PIAAC (2012), where the ability to employ the digital tools provided in simulated web, email, and spreadsheet environments is integral to the construct. We can expect that other examples will emerge as different technologies become increasingly embedded in everyday life.

As the number of measured constructs increases in conjunction with the possible introduction of rotated background questionnaire designs, the stability of the key latent regression model may be reduced. This, in turn, would threaten the accuracy of the population model used to generate PVs—especially in the case of sparse designs for the cognitive instruments. Present concerns regarding the impact on proficiency estimates of measurement error in the explanatory variables in the latent regression model will likely become more pronounced (L. Rutkowski & Zhou, 2015). Presumably, at some point there will have to be some trade-offs between relevance and maintenance of the reputation of ILSAs as trustworthy sources of achievement data. Other trade-offs between local fidelity and international comparability will be informed by solid analytic work but will also necessarily reflect policy and political considerations. Consequently, the decisions reached may place additional burdens on the ILSA teams, which will have to accommodate greater variation among the instruments administered in different countries.

Another critical aspect of relevance concerns the degree of alignment between the information provided by the ILSAs and the questions asked by policy makers and other stakeholders. In point of fact, a number of key issues that command some degree of general interest are frequently raised. To improve timeliness, the relevant data could be culled from the comprehensive almanacs and organized and displayed relatively quickly in tables and figures. Further on, targeted secondary analyses could be funded through sponsored initiatives—perhaps initiated even before the public release of the data.

Relevance could be increased by more systematically linking national ILSA data to national databases that contain other types of information (e.g., background data at the individual level). Such an augmented database would support a broader range of secondary analyses, as noted previously in the case of Denmark. Linkages to data at higher levels of aggregation (e.g., at the school or area level) could also prove useful for some analyses. Going forward, designers could build in some connections as a national option to facilitate such linkages.

POLICY AND POLITICAL CHALLENGES

Alongside the technical concerns that relate primarily to construct validity, we now turn to a discussion of some of the political challenges that can be framed in terms of consequential validity. The increased prominence of ILSAs results in policy discussions, as well as the exhortations on the part of some leaders that countries emulate the “high-flyers,” has sparked a backlash. Some researchers and commentators have argued that ILSA results have become too influential in national policy discussions and, in particular, that they are a force for “international homogenization” at the expense of national educational differences that should be preserved (Carnoy, 2015; Grek, 2009; Meyer & Benavot, 2013; Sellar & Lingard, 2013). They cited the broader participation in ILSAs, the increased salience of ILSA results in discussions of education policy, and the utilization of ILSA outcomes for purposes of evaluation and monitoring.

Benavot and Smith (2020) linked increasing participation in ILSAs in part to efforts by the UNESCO Institute for Statistics to develop global learning metrics as a means for quantifying progress toward the United Nations’ Sustainable Development Goal 4: Education. By participating in an international comparative assessment, countries can satisfy the SDG 4 reporting requirements. The authors argue that these requirements have also led to pressure to conform to a system of “global educational governance.” Teacher associations have also weighed in on ILSA participation (Couture, 2016).

A review of the literature, however, suggests that the situation is rather more ambiguous and that a more nuanced view is in order. Successful policy transfer across national boundaries can be difficult—for both technical and political reasons (Atkin & Black, 1997; Braun, 2008; Burdett & O’Donnell, 2016; Oliveri et al., 2018). Drawing on extensive personal experience, Ritzen (2013) observed that the degree of utilization of ILSA results depends not only on a country’s readiness (politically and otherwise) to institute changes, but also on its capacity and commitment to carry out the change process. This observation may be especially true for many low- and middle-income countries, whose participation has been essentially mandated (and supported by) international donor organizations (see also Braun, 2013).

Volante et al. (2017), in discussing the normative impact of the assessments sponsored by the OECD, also noted that there is a wide range of policy responses to ILSA results—from negligible to modest to very substantial. Fischman et al. (2019) conducted surveys of ILSA experts and interested stakeholders. They found considerable diversity of opinion regarding whether ILSAs had an impact on national education policymaking, and if so, whether it was constructive. The observed variation is likely due both to differences in experiences among countries and to differences in the vantage points of the respondents.

These debates can be framed in terms of questions regarding the consequential validity of ILSAs. That is, beyond the attribution of impact (see the section “Impact”), the principal issue lies in what cases the impact leads to long-term positive outcomes and

in what cases it does not—and whether the two can be reasonably well distinguished. This section addresses the issue.

Concerns Regarding Impact

Once a country joins an ILSA collaborative, an obvious benefit of continued participation is the value of tracking performance trends over time. It is impossible to ignore the “league tables” that are published after each ILSA data release. However, a focus on rankings based on aggregate mean scores, or changes in rankings over time, can be misleading and lead to confusion among stakeholders. First, countries with mean scores that are statistically close nonetheless could have quite different rankings. Second, many factors determine a country’s rank in a particular cycle (e.g., the set of countries participating in that cycle) that have little or nothing to do with the efficacy of its education system (Singer & Braun, 2018).

Nonetheless, it is true that some ILSA proponents argue that a country’s best strategy for educational improvement is to emulate the policies of so-called high flyers (National Center for Education and the Economy, 2020; Schleicher, 2018). Kamens (2013) argued that this sort of cheerleading encourages stakeholders to think that there is a “magic bullet” to achieving success. The problem is sometimes exacerbated by misleading media reports that may focus on changes in a country’s rank but ignore the actual change in performance. PISA 2015 offers some examples from East Asia. In Japan, one newspaper highlighted a change in rank from fourth to second on science, even though the mean score declined from 547 to 538. By contrast, in Chinese Taipei (Taiwan), a newspaper bemoaned a decline in ranking on reading, even though the mean score increased.⁵⁸ To counter these and other misinterpretations, there have been some innovative attempts to present country results through visuals that de-emphasize league tables.

Some authors have argued that for the United States, state-level comparisons are likely to be more informative than international comparisons (Carnoy et al., 2015).⁵⁹ For now, that appears to be a minority view. Indeed, these same authors suggested that cross-country comparisons among recognizable subgroups are more informative than aggregate comparisons. For example, they compared countries’ performance in mathematics on PISA 2012 within strata defined by a cross-national measure of family academic resources. They showed that although the United States fares poorly overall, it does relatively well in each stratum. Its comparatively lower aggregate scores are due to the greater proportions of U.S. students in the more disadvantaged strata. This finding leads to consideration of different reform strategies than might be suggested by a singular focus on overall rankings. Similar studies based on data from PIAAC could also yield interesting patterns.

Another widespread concern is related to misinterpretations and misuses that are not directly due to technical deficiencies of the assessment or the failure to provide appropriate guidelines (Kane, 2016). That is, despite the best efforts of sponsors and developers, either policy makers make inappropriate policy pronouncements or decisions or various entities overstate the evidential value of ILSA results with respect to policy and

practice. For example, in some instances, national education goals have been framed in terms of specific improvements in country rankings. D. Rutkowski et al. (2020) cited the example of Australia. The Education Act of 2012, approved by the Parliament of the Commonwealth of Australia, called for Australia to be ranked by PISA 2025 in the “top five” on reading, mathematics, and science. A misguided focus on the rankings competition can lead to suboptimal policy choices.

On the one hand, although they are not directly responsible for such misuses, it is still incumbent on ILSA sponsors and developers to counter specific misuses and misinterpretations of ILSA data. On the other hand, a call to reduce the gaps in scores between high and low performers within a country to a level comparable to the gaps seen in peers with comparable overall performance but greater equity should be seen as a constructive use of ILSA data for purposes of benchmarking.

Responding to Political Challenges

Judgments regarding whether ILSAs exert undue influence on national education policies should take into account the country’s rationale for participation. As noted in the section “Impact,” ILSA results are sometimes used to justify or legitimate pending policies. Wiseman (2013) offered the example of France. Citing Dobbins and Martens (2012), he stated that “the OECD’s PISA was part of a French political agenda to create an evidence base to support a specific policy position. This is not an isolated tactic or unusual use of international achievement studies” (Wiseman, 2013, p. 311). J. Jennings noted that in the United States, the National Council of State Legislators in 2016 seized on the then–most recent PISA results to build support for a wide range of education reforms.⁶⁰ In Ireland, surprisingly weak PISA results provided impetus for the passage of long-stalled education reforms (H. Hislop, personal communication, September 21, 2017). Asserting that these represent constructive uses of ILSA data depends largely on a judgment of the appropriateness of the reform policies.

Wiseman (2013) also argued that, to the extent that ILSAs exert influence on educational policies, the outcome is not necessarily some form of strict policy convergence. Instead, the result may be an “isomorph” of the “model” system, whereby countries extract some key features of the model (e.g., curriculum) but then modify and/or adapt them to better fit the country’s traditions and culture, often taking into account the political strengths of different stakeholders. Thus, despite exhortations in support of emulation, the path from ILSA results to policy responses is neither straightforward nor predictable (Pons, 2017).

Ideally, a country’s decision to participate in a particular, school-based ILSA would be informed by careful study of the ILSA’s avowed purposes and its assessment frameworks, followed by an evaluative judgment regarding their degree of congruence with the country’s goals for its students (Oliveri et al., 2018). Presumably, the reputation of the ILSA for generating data that are accurate, reliable, and valid would also factor into the decision. In this context, careful study of a country’s outcomes could very well appropriately result in changes in curriculum and pedagogy. As described in the section “Impact,” this strategy has been adopted by many countries—in some cases

(e.g., Singapore) in a thoughtful, systematic fashion. Despite the caveats raised in that section, the strategy may be particularly helpful to countries with lower capacity in the realms of curricular design, test development, and educational measurement. This sort of “homogenization” is consistent with Wiseman’s (2013) model of convergence.

It bears repeating that country differences in ILSA-based indicators cannot be directly linked to differences in the quality of their education systems (Singer et al., 2018). More nuanced contextual analysis is in order. For example, the strong performances of students in South Korea and Japan are due, in part, to the pervasive practice of out-of-school tutoring (shadow education), as noted by Heyneman (2013) and others. Countries also differ in their investments in both the well-being of children and teacher quality. These and other differences can contribute to differences in academic achievement, particularly for children living in less advantaged circumstances. A failure to appreciate this complexity can lead to misinterpretations of ILSA results (particularly changes in overall rankings) and, hence, to misguided policy decisions. It is worth noting that such differences—and their misinterpretations—can also occur when making comparisons across population subgroups or geographical regions within a country.

Despite the appropriate concerns regarding the overemphasis on league tables, rankings on other measures can provide useful policy information. For example, one country-level indicator of equity is the difference in the mean scores between the top decile (quintile) and the bottom decile (quintile), with larger differences signaling greater inequality. A related indicator is the slope of the regression of cognitive scores on a measure of socioeconomic status. Again, a steeper slope signals greater inequality. It would be useful to consider disaggregating cognitive outcomes data by other background variables to provide more informative comparisons (Rowley et al., 2020). In school-based ILSAs, for example, reporting results by levels of opportunity to learn (or available proxies) should be considered. In reporting results to their national policy makers, countries should consider presenting a range of indicators. It is particularly striking when countries with similar overall proficiency means differ substantially on one or the other of these indicators. Such reports must include clear, explanatory text to highlight the utility of these indicators and to reduce the possibility of misinterpretations.

To highlight the benefits of participation in school-based ILSAs, sponsors, in conjunction with country representatives and education authorities, should enhance outreach to teachers, curriculum specialists, and education leaders to explicate the ILSA frameworks and how the assessments reflect those frameworks. They could also organize specially designed workshops or webinars that highlight important issues or findings that might help guide interpretations. With a solid understanding of how to interpret the outcomes, the classroom utility of both the frameworks and the disaggregated results would be more evident. Country representatives who participate in the various expert groups during an ILSA cycle are an underutilized resource. They not only help to define and operationalize the various frameworks, but also contribute to the development of described proficiency scales. Thus, they are well positioned to inform education policy discussions within a country.

Similar considerations are relevant to participation in adult surveys. Outreach to various government ministries and other stakeholders could emphasize the value of the data collected by PIAAC for policy decisions related to a range of domains, including education, training, and labor market initiatives. Moreover, the argument can be strengthened by reference to the many secondary analyses already conducted on available data.⁶¹ The results of these analyses enable useful cross-national comparisons that can directly inform policy discussions.

In general, the “duty of care” incumbent on ILSA sponsors and developers to address concerns regarding consequential validity can be framed in terms of ensuring, to the extent possible, that (a) the technical quality of the results supports their intended interpretations and uses; (b) the presentation of the results, along with the corresponding explanatory materials, encourages appropriate interpretations and uses; (c) the text is transparent about possible biases and uncertainties, as well as their implications for interpretation; and (d) the most common, anticipated misinterpretations and misuses are proactively addressed and discouraged. This aspect of validity is all the more important with the increases in ILSA complexity and heterogeneity of participating countries.

CONCLUSION

The inclusion of this chapter in the present volume reflects the growing importance of ILSAs of student and adult populations since the 1960s. From modest beginnings as a set of exploratory studies devised by a small group of scholars interested in international comparative education, school-based ILSAs have grown into a major strand of applied research within the larger domain of comparative education (Suter et al., 2019). As the scope and reach of ILSAs have expanded, they have generated an unprecedented amount of empirical evidence on which researchers and policy makers can draw to analyze education systems and develop strategies to improve student outcomes. Nonetheless, some worry that by virtue of their dominant role, ILSAs have led to a diminished appreciation for the insights that other strands of comparative education have to offer—not the least of which is providing a broader framework within which ILSA findings can be understood and applied (Carnoy, 2019). Indeed, as is evident in the contributions to Suter et al. (2019), these other strands rest on foundations of both theory and rich, empirical studies.

At the same time, the increased attention paid to ILSA scores and rankings, as well as to the results of secondary analyses of these data, is due in large measure to the widespread recognition of the importance of understanding the relationships of education and skills, not only to social and economic development, but also to general well-being.⁶² As globalization and technology have accelerated, the economic interdependencies (and competition) among countries, policy makers, and key stakeholders have seen value in benchmarking performance against peers and near-peers—initially in more developed economies and, more recently, in middle- and lower income countries.

In 1987, following the implementation of the new NAEP design, Messick published a paper in which he asserted that because large-scale assessments constitute a form of policy research, their success should be judged based on their policy utility. Further, he reasoned that an element of uncertainty exists between policy research and policy formation, creating a gap that requires informed judgment. He believed that large-scale assessments could provide information to help bridge this gap through the provision of evidence gathered by careful design and development of instruments, followed by appropriate analysis and reporting of results (Kirsch & Braun, 2020; Lennon & Kirsch, 2025). He did not assert, however, that this empirical information was sufficient to the purpose.

To help understand how large-scale assessments could best fulfill this role, Messick (1987) proposed a framework containing three key design criteria: comparability, interpretability, and relevance. This chapter has presented and extended these three design criteria, arguing that they remain applicable for evaluating the utility of next-generation ILSAs. Indeed, a recurring theme in the chapter has been how changes to ILSAs have contributed to greater utility by strengthening one or more of the design criteria.

For example, to enhance relevance, ILSAs are responding to broad interest in new, more innovative domains such as creative thinking, problem-solving, and others that are most suitable for digitally based assessments. However, as always, there are trade-offs. In this case, countries or individuals who receive the paper-based assessments do not participate in these innovative domains, potentially reducing the utility of these surveys for these countries and groups. Moreover, the evolution of ILSAs in this direction will result in a growing divergence between the two modes in what is being measured and, consequently, weakening the justification for reporting results on a single scale. Unfortunately, both practical and political considerations argue for maintaining two delivery modes, with the attendant complications and additional costs.

In addition to employing Messick's (1987) framework to examine various aspects of ILSA development and subsequent uses, this chapter also has described the ILSA developmental path as a virtuous spiral, that is, a trajectory in which each cycle of an ILSA attempts to meet the evolving interests of policy makers and key stakeholders who continually challenge sponsors and researchers to generate information that is relevant to the changes taking place within and across participating countries. As societies undergo changes, policy makers and other stakeholders pose new questions that lead to both novel and better measures of legacy constructs and the introduction of new domains of assessment.

In this context, we highlighted two key inflection points along this trajectory that have led to marked increases in ILSA utility and salience. The first inflection point occurred in the mid-1980s through the mid-1990s with the introduction of important methodological innovations such as BIB spiraling and IRT scaling. These innovations provided policy makers with interpretable information regarding the distributions of skills within and across countries, along with the ability to examine

the relationships among these skills and a host of demographic and background variables.

Over the next 2 decades, additional methodological and technological innovations ultimately led to the second inflection point, one characterized by the advent of digital platforms that support the design, development, and delivery of next-generation ILSAs. Successfully carrying out digitally-based assessments required the construction of novel tools and the integration of new workflows and processes that benefited from the ongoing collaboration of experts across the different phases. The digital platform facilitated the introduction of new constructs, new item types, and improved methodologies, such as adaptive multistage testing. The result has been an improvement in the overall quality of the data, extensions of what can be measured in both student and adult populations, and greater efficiency in the management, delivery, and processing of the data.

As ILSAs have proven over the years to be a reliable source of credible information, this chapter also described some of the ways in which these assessments have impacted policy. Examples include countries where new policies were implemented based on results from in-school surveys including PIRLS, TIMSS, and PISA. There are other cases where these assessments were used to support existing policies or legitimate proposed policies. Beyond their direct impact, there is also evidence of ILSAs' indirect impact, which is reflected in increased media attention and the documented growth in the number of researchers who are using ILSA databases to inform their research and develop policy papers covering a range of issues relevant to school-age and adult populations. Finally, ILSA participation has jump-started or accelerated capacity building, particularly in many middle- and low-income countries, giving them the capability to conduct national or subnational assessments to inform education policy.

Not surprisingly, the growth in relevance and popularity of ILSAs has also led to a number of challenges and concerns. One challenge arises from the substantial increase in the diversity of participating countries, as well as that of social and economic backgrounds, representing a wider range of skills. Maintaining an adequate level of measurement invariance across countries among both the cognitive domains and the constructs underlying the background questionnaires has become increasingly difficult as participation has expanded.

Compounding the challenges presented by the growing diversity among participating countries is the never-ending pressure to extend what is being assessed. Because technology-based assessments facilitate such extensions, there is continued demand to add to the constructs in the background questionnaire, as well as those in the cognitive domains. Given the constraints on survey administration time, these extensions necessitate the development and deployment of more complex designs, such as within-construct rotation of key variables in the background questionnaire—similar to what is done in the cognitive domains.

Predictably, increases in ILSA scope and coverage lead to rising costs for both contractors and participating countries. Greater operational difficulties due to lower

response rates, as well as to the hiring, training, and retaining of survey staff, also contribute to rising costs. These increases cause concerns among the governments who not only fund the assessments, but also must wait patiently for several years from the initiation of survey development to the final reporting of results.

More recently, the questions that have been raised are of a more political nature and have less to do with technical issues. For example, there are widespread concerns regarding the overemphasis on the league tables of country rankings and their relative changes over time. This attention to league tables is due in part to the (misguided) notion that differences in countries' rankings are a credible indicator of differences in the efficacies of their education systems. A natural conclusion, then, is that a sure-fire strategy recognized for educational improvement involves trying to identify and implement those policies that seem to be effective in high-performing countries (Tucker, 2011).

Aside from technical issues related to data quality (as was the case with Shanghai),⁶³ the difficulty with this advice is that it fails to acknowledge the natural limitations of ILSA data; namely, they can suggest interesting hypotheses but cannot establish them, largely, but not solely, because of the constraints on making causal inferences from cross-sectional data. More intensive studies are needed, requiring consideration of the broad range of factors operating at different levels from conception to the time of the assessment that contribute to student achievement. It is precisely here that the findings from "traditional" comparative education come into their own. Understanding the historical, cultural, economic, and political influences on the structure and functioning of current education systems—and how they differ across countries—is essential to devising workable policies and interventions that mitigate some of the likely difficulties that arise in a policy transfer process.

As noted in the last section, consideration of ILSA impact (both positive and negative) falls under the rubric of consequential validity. The responsibilities associated with consequential validity lie with both the sponsors and the developers, as well as with those who use and interpret the data. The sponsors and developers are expected to, and often do, provide information regarding the overall quality of the data, along with the appropriate uses and interpretations of the results. The media, policy makers, and others are expected to assume some responsibility for misinterpretations or overstatements of what the data say or what decisions can be supported by the results.

These challenges and concerns, both technical and political in nature, constitute healthy tensions that can and should be discussed and debated within the extended Messick (1987) framework. Although it is certain that ILSAs will continue to face a number of challenges, we believe that the path along the virtuous spiral will continue to provide innovations that will enable ILSAs to address new and more complex questions. We expect that answers to these questions will lead to continued appreciation for the utility for such surveys.

ACKNOWLEDGMENTS

The authors wish to thank the editors, L. Cook and M. Pitoniak, for inviting them to contribute a chapter to the volume and for their support throughout the process. Reviewers of our first draft, A. Ben Simon, M. von Davier, and H. Wagemaker, offered a number of useful comments and suggestions that we endeavored to implement. Subsequent drafts were reviewed critically by M. L. Lennon and E. Gonzalez. Both devoted substantial time to the reviews and we benefited tremendously from their many suggestions, almost all of which we were able to follow. On specific matters we were assisted by B. Fishbein, P. Foy, M. O. Martin, L. Suter, and K. Yamamoto. Finally, A. K. Gontz was absolutely thorough in conducting the final editing of the manuscript. To all we are very, very grateful.

REFERENCES

- Ainley, M., & Ainley, J. (2019). Non-cognitive attributes: Measurement and meaning. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 103–125). SAGE Publications. <https://doi.org/10.4135/9781526470379.N7>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Atkin, J. M., & Black, P. (1997). Policy perils of international assessments. *Phi Delta Kappan*, 79(1), 22–28.
- Avvisati, F. (2020). The measure of socio-economic status in PISA: A review and some suggested improvements. *Large-Scale Assessments in Education*, 8, Article 8. <https://doi.org/10.1186/s40536-020-00086-x>
- Baird, J.-A., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research*, 58(2), 121–138. <https://doi.org/10.1080/00131881.2016.1165410>
- Beaton, A. E., & Barone, J. L. (2017). Large-scale group score assessments. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 233–284). Springer Open. https://link.springer.com/chapter/10.1007/978-3-319-58689-2_8
- Bedard, K., & Dhuey, E. (2006). The persistence of early childhood maturity: International evidence of long-run age effects. *The Quarterly Journal of Economics*, 121(4), 1437–1472. <https://doi.org/10.1093/qje/121.4.1437>
- Benavot, A., & Smith, W. (2020). Reshaping quality and equity: Global learning metrics as a ready-made solution to a manufactured crisis. In A. Wulff (Ed.), *Grading goal four: A strategic take on the tensions, threats and opportunities in the sustainable development goal on quality education* (pp. 238–261). Brill. https://www.research.ed.ac.uk/files/154046014/Reshaping_quality_and_equity_GLMs_as_a_ready_made_solution_Benavot_Smith_2020.pdf

- Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44(6), 706–732. <https://doi.org/10.3102/1076998618784700>
- Biemer, P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lysberg, L. E., Tucker, N. C., & West, B. T. (Eds.). (2017). *Total survey error in practice*. John Wiley & Sons. <https://doi.org/10.1002/9781119041702>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer. https://doi.org/10.1007/978-94-007-2324-5_2
- Bracey, G. W. (2008, Fall). Disastrous legacy: Aftermath of a nation at risk. *Dissent Magazine*. <https://www.dissentmagazine.org/article/disastrous-legacy-aftermath-of-a-nation-at-risk>
- Bradburn, N. M., & Gilford, D. M. (Eds.). (1990). *A framework and principles for international comparative studies in education*. National Academies Press. <https://doi.org/10.17226/9220>
- Braun, H. (2008). Review of McKinsey report: How the world's best performing school systems come out on top. *Journal of Educational Change*, 9(3), 317–320. <https://doi.org/10.1007/s10833-008-9075-9>
- Braun, H. (2013). Prospects for the future: A framework and discussion of directions for the next generation of international large-scale assessments. In M. von Davier, I. Kirsch, K. Yamamoto, & E. Gonzalez (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy and educational research* (pp. 149–160). Springer Science and Business Media. https://doi.org/10.1007/978-94-007-4629-9_8
- Braun, H. (2018). How long is the shadow? The relationships of family background to selected adult outcomes: Results from PIAAC. *Large-Scale Assessments in Education*, 6, Article 4. <https://doi.org/10.1186/s40536-018-0058-x>
- Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record*, 113(11), 2309–2344. <https://doi.org/10.1177/016146811111301101>
- Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring educational systems: International comparisons. *Annals of the Academy of Political and Social Sciences*, 683(1), 75–92. <https://doi.org/10.1177%2F0002716219843804>
- Braun, H., & von Davier, M. (2018). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-Scale Assessments in Education*, 5, Article 17. <https://doi.org/10.1186/s40536-017-0050-x>
- Braun, H. I., Wang, A., Jenkins, F., & Weinbaum, E. (2006). The Black–White achievement gap: Do state policies matter? *Education Policy Analysis Archives*, 14, Article 8. <https://doi.org/10.14507/epaa.v14n8.2006>
- Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers, No. 71). OECD. <https://doi.org/10.1787/5k9fdfqfrr28-en>

- Burdett, N., & O'Donnell, S. (2016). Lost in translation? The challenges of educational policy borrowing. *Educational Research*, 58(2), 113–120. <https://doi.org/10.1080/00131881.2016.1168678>
- Byrne, B., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema*, 29(4), 539–551. <https://doi.org/10.7334/psicothema2017.178>
- Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). Chapman & Hall/CRC.
- Carnoy, M. (2015). *International test score comparisons and educational policy: A review of the critiques*. National Education Policy Center. <https://nepc.colorado.edu/publication/international-test-scores>
- Carnoy, M. (2019). The uneasy relation between international testing and comparative education research. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 569–595). SAGE Publications. <https://doi.org/10.4135/9781526470379.n31>
- Carnoy, M., Garcia, E., & Kavenson, T. (2015). Bringing it back home: Why state comparisons are more useful than international comparisons for improving U.S. education policy (Briefing Paper No. 410). Economic Policy Institute. <https://www.epi.org/publication/bringing-it-back-home-why-state-comparisons-are-more-useful-than-international-comparisons-for-improving-u-s-education-policy/>
- Clarke, M., & Luna-Bazaldua, D. (2021). *Primer on large-scale assessments of educational achievement*. World Bank. <https://doi.org/10.1596/978-1-4648-1659-8>
- Conaway, C. (2020). Maximizing research use in the world we actually live in: Relationships, organizations, and interpretation. *Education Finance and Policy*, 15(1), 1–10. https://doi.org/10.1162/edfp_a_00299
- Couture, J.-C. (2016, May 31). Association to push for PISA withdrawal. *ATA News*, 50(18). The Alberta Teachers' Association. <https://legacy.teachers.ab.ca/News%20Room/ata%20news/Volume%2050%202015-16/Number-18/Pages/PISA-withdrawal.aspx>
- Cunningham, S. (2021). *Causal inference: The mixtape*. Yale University Press. <https://doi.org/10.12987/9780300255881>
- Darling-Hammond, L. (2012). Policy frameworks for new assessments. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 301–339). Springer Open. https://doi.org/10.1007/978-94-007-2324-5_6
- Dept, S., Ferrari, A., Souto Pico, M., & Lupsa, D. (2025). Translation in Computer-based international large scale assessment: Not a stand-alone component. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments—foundations, methodologies, and quality assurance procedures* (pp. 191–219). Springer Open.
- Dobbins, M., & Martens, K. (2012). Towards an education approach à la finlandaise? French education policy after PISA. *Journal of Education Policy*, 27(1), 23–43. <https://doi.org/10.1080/02680939.2011.622413>
- Durrant, G., & Schnepf, S. (2017). Which schools and pupils respond to educational achievement surveys? A focus on the English Programme for International Student

- Assessment sample. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 181(4), 1057–1075. <https://doi.org/10.1111/rssa.12337>
- Ebbs, D., & Wry, E. (2016). Translation and layout verification for PIRLS 2016. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 1–15). TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-7.html>
- Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning using examinee response processes for the next generation of assessments*. Routledge. <https://doi.org/10.4324/9781315708591>
- Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, 32(5), 619–634. <https://www.jstor.org/stable/4618685>
- Feuer, M. J. (2012). *No country left behind: Notes on the rhetoric of international comparisons of education* (William Angoff Invited Lecture). ETS.
- Feuer, M. J. (2013). Validity of international large-scale assessments: “Truth and consequences.” In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 197–216). Emerald Group Publishing.
- Fischman, G., Marcetti Topper, A., Silova, I., Goebel, J., & Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy*, 34(4), 470–499. <http://doi.org/10.1080/02680939.2018.1460493>
- Fishbein, B., Foy, P., & Tyack, L. (2020). Chapter 10: Reviewing the TIMSS 2019 achievement item statistics. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 10.1–10.70). TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods/chapter-10.html>
- Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assessments in Education*, 6, Article 11. <https://doi.org/10.1186/s40536-018-0064-z>
- Fogg, N., Harrington, P., & Khatiwada, I. (2018). *Skills and earnings in the full-time labor market*. ETS. <https://www.ets.org/s/research/pdf/skills-and-earnings-in-the-part-time-labor-market.pdf>
- Glas, C., & Jehangir, K. (2013). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 97–115). CRC Press.
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). OECD Publishing. <https://doi.org/10.1787/5jlzfl6fhxs2-en>
- Goldstein, H. (2004). International comparative assessment: How far have we really come? *Assessment in Education*, 11(2), 227–234. <https://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/intl-comp-assessment-rev-essay.pdf>

- Goodman, M. J., Sands, A. M., & Coley, R. J. (2015). *America's skills challenge: Millennials and the future*. ETS. <https://www.ets.org/s/research/30079/asc-millennials-and-the-future.pdf>
- Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Education Policy*, 24(1), 23–37. <https://doi.org/10.1080/02680930802412669>
- Grund, S., Ludtke, O., & Robitzsch, A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, 46(4), 430–465. <https://doi.org/10.3102/1076998620959058>
- Gurkan, G. (2021). *From OLS to multilevel multidimensional mixture IRT: A model refinement approach to investigating patterns of relationships in PISA 2012 data* [Doctoral dissertation, Boston College]. <http://hdl.handle.net/2345/bc-ir:109191>.
- Gustafsson, J. E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, 24(3), 275–295. <https://doi.org/10.1080/09243453.2013.806334>
- Gustafsson, J. E., & Nilson, T. (2016). The impact of school climate and teacher quality on mathematics achievement: A difference-in-differences approach. In T. Nilson & J. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts, and time* (Vol. 2, pp. 81–95). Springer. https://link.springer.com/chapter/10.1007/978-3-319-41252-8_4
- Haertel, E. H. (2018). Tests, test scores, and constructs. *Educational Psychologist*, 53(3), 203–216. <https://doi.org/10.1080/00461520.2018.1476868>
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory: Vol. 1. Methods and applications*. John Wiley.
- Hanushek, E. A., Schwerdt, G., Wiederhold, S., & Woessmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review*, 73, 103–130. <https://doi.org/10.1016/j.eurocorev.2014.10.006>
- He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, 26(4), 369–385. <https://doi.org/10.1080/0969594X.2018.1469467>
- Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education*, 9, Article 17. <https://doi.org/10.1186/s40536-021-00109-1>
- Heyneman, S. P. (2013). The international efficiency of American education: The bad and the not-so-bad news. In H. Meyer & A. Benavot (Eds.), *PISA, power, and policy. The emergence of global educational governance* (pp. 279–302). Symposium Books.
- Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37–72). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-7/impact->

international-studies-academic-achievement-policy-research-stephen-heyneman-bommi-lee?context=ubx&refId=f06897f6-cdc8-40d9-bfd6-824b5d9b9b34

- Hooper, M. (2017). *Applying the pseudo-panel approach to international large-scale assessments: A methodology for analyzing subpopulation trend data* [Doctoral dissertation, Boston College]. <https://eric.ed.gov/?id=ED578771>
- Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of 12 countries* (Vol. 2). John Wiley.
- International Association for the Evaluation of Educational Achievement. (n.d.-a). *ILSA Gateway*. lsa-gateway.org
- International Association for the Evaluation of Educational Achievement. (n.d.-b). *Tools*. <https://www.iea.nl/data-tools/tools>
- International Association for the Evaluation of Educational Achievement. (2021). *Help manual for the IEA IDB analyzer* (Version 5.0).
- Jones, L. V., & Olkin, I. (Eds.). (2004). *The nation's report card: Evolution and perspectives*. Phi Delta Kappa Educational Foundation.
- Kamens, D. H. (2013). Globalization and the emergence of an audit culture: PISA and the search for “best practices” and magic bullets. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance* (pp. 117–140). Symposium Books.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedem.12000>
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198–211. <https://doi.org/10.1080/0969594X.2015.1060192>
- Kane, M. T., Berryman, S., Goslin, D., & Meltzer, A. (1990). *The secretary's commission on achieving necessary skills: Identifying and describing the skills required by work*. Pelavan Associates. <https://wdr.doleta.gov/scans/idsrw/idsrw.pdf>
- Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large-scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-Scale Assessments in Education*, 1, Article 6. <https://doi.org/10.1186/2196-0739-1-6>
- Karabenick, S. A., Wooley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C., De Groot, E., Gilbert, M. C., Musu, L., Rogat, T., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, 42(3), 139–151. <https://doi.org/10.1080/00461520701416231>
- Kirsch, I. S. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (ETS Research Report No. RR-01-25). ETS. <https://doi.org/10.1002/j.2333-8504.2001.tb01867.x>
- Kirsch, I. S. (2003). Measuring literacy in IALS: A construct-centered approach. *International Journal of Educational Research*, 39(3), 181–190. <https://doi.org/10.1016/j.ijer.2004.04.002>
- Kirsch, I. S., & Braun, H. I. (2020). Changing times, changing needs: Enhancing the utility of international large-scale assessments. *Large-Scale Assessments in Education*, 8, Article 10. <https://doi.org/10.1186/s40536-020-00088-9>

- Kirsch, I. S., Braun, H., Lennon, M. L., & Sands, A. (2016). *Choosing our future: A story about opportunity in America*. ETS. <https://www.ets.org/s/research/report/opportunity/ets-choosing-our-future.pdf>
- Kirsch, I. S., Braun, H., Yamamoto, K., & Sum, A. (2007, January). *America's perfect storm* [Policy Information Report]. ETS. <https://www.ets.org/Media/Research/pdf/PICSTORM.pdf>
- Kirsch, I. S., Lennon, M. L., Yamamoto, K., & von Davier, M. (2017). Large-scale assessments of adult literacy. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 285-310). Springer Open. https://doi.org/10.1007/978-3-319-58689-2_9
- Kroehne, U., Roelke, H., Kuger, S., Goldhammer, F., & Klieme, E. (2016, April 7–11). *Theoretical framework for log-data in technology-based assessments with empirical applications from PISA* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Washington, DC, United States.
- Kuger, S., Kleime, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. Springer. <https://doi.org/10.1007/978-3-319-45357-6>
- Kyllonen, P. C., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment data analysis: Background, technical issues, and methods of data analysis* (pp. 277–286). Chapman & Hall/CRC Press. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-17/innovative-questionnaire-assessment-methods-increase-cross-country-comparability-patrick-kyllonen-jonas-bertling?context=ubx&refId=2eff3ab3-f0ae-4957-abb5-ad70d13271d9>
- Laitusis, C. C., King, T., Cavalie, C., James, K., & Finnegan, A. (2018). *PISA special education needs feasibility study*. OECD.
- Lennon, M. L., & Kirsch, I. (2025). Innovations in item development for computer-based assessments. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments—foundations, methodologies and quality assurance procedures* (pp. 21–64). Springer Open.
- Leu, D. J., Kiili, C., Forzani, E., Zawilinski, L., O'Byrne, W. I., & McVerry, J. G. (2021). New literacies of online research and comprehension. In C.A. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 844–852). Wiley–Blackwell. <https://doi.org/10.1002/9781405198431.wbeal0865.pub2>
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.
- Lockheed, M. (2013). Causes and consequences of international assessments in developing countries. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power and policy: The emergence of global educational governance* (pp. 163–183). Symposium Books. <https://doi.org/10.15730/books.85>
- Lohr, S. L. (2010). *Sampling: Design and analysis*. Brooks/Cole.
- Loveless, T. (2014). *The 2014 Brown Center report on American education: How well are American students learning?* The Brookings Institution. <https://www.brookings>

edu/wp-content/uploads/2016/06/2014-Brown-Center-Report_FINAL-4.pdf

- Luo, X. (2020). Automated test assembly with mixed-integer programming: The effects of modeling approaches and solvers. *Journal of Educational Measurement*, 57(4), 547–565. <https://doi.org/10.1111/jedm.12262>
- Maehler, D. B., Jakowatz, S., & Konradt, I. (2020). *PIAAC Bibliography—2008–2019* (GESIS Papers, 2020/04). GESIS, Leibniz-Institut für Sozialwissenschaften. <https://doi.org/10.21241/ssoar.67732>
- Marchionni, M., & Vazquez, E. (2019). The causal effect of an extra year of schooling on skills and knowledge in Latin America: Evidence from PISA. *Assessment in Education: Principles, Policy & Practice*, 26(4), 489–515. <https://doi.org/10.1080/0969594X.2018.1454401>
- Marks, G. N., & O’Connell, M. (2021). Inadequacies in the SES—achievement model: Evidence from PISA and other studies. *Review of Education*, 9(3), e3293. <https://doi.org/10.1002/rev3.3293>
- Martin, M. O., Foy, P., Mullis, I. V., & O’Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at fourth grade. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—implications for early learning* (pp. 109–180). TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf
- Martin, M. O., Mullis, I. V. S., Arora, A., & Preuschoff, C. (2014). Context questionnaire scales in TIMSS and PIRLS 2011. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-19/context-questionnaire-scales-timss-pirls-2011-michael-martin-ina-mullis-alka-arora-corinna-preuschoff?context=ubx&refId=ab5d667f-8fe2-4905-bebf-fb86a500a3b6>
- Martin, M. O., Mullis, I. V. S. & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/publications/timss/2015-methods.html>
- Martin, M. O., Mullis, I. V. S., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/publications/pirls/2016-methods.html>
- Martin, M. O., Rust, K., & Adams, R. (Eds.). (1999). *Technical standards for IEA studies*. International Association for the Evaluation of Educational Achievement. <https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies>
- Martin, M. O., von Davier, M., & Mullis, I. V. S. (2020). *Methods and procedures: TIMSS 2019 technical report*. TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/timss2019/methods/index.html>
- Mazzeo, J., Lazer, S., & Zieky, M. J. (2006). Monitoring educational progress with group-score assessments. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 681–699). American Council on Education/Praeger.

- Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229-258). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-15/linking-scales-international-large-scale-assessments-john-mazzeo-matthias-von-davier?context=ubx&refId=b641e638-90ba-40b2-8991-29db10e0ed5a>
- Meinck, S., Gonzalez, E., & Wagemaker, H. (2020). Consequential validity: Data access, data use, analytical support, and training. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 231-244). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-030-53081-5_13
- Messick, S. J. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. *European Journal of Psychology of Education*, 2, Article 157. <https://doi.org/10.1007/BF03172645>
- Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on Education/Macmillan Publishing Company.
- Messick, S. J. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23. <https://doi.org/10.3102/0013189X023002013>
- Messick, S. J., Beaton, A. E., & Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Report No. 83-1). ETS.
- Meyer, H.-D., & Benavot, A. (2013). PISA and the globalization of education governance: Some puzzles and problems. In H.-D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 9-26). Symposium Books. <https://doi.org/10.15730/books.85>
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177-196. <https://doi.org/10.1007/BF02294457>
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. <https://doi.org/10.4324/9781315871691>
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161. <https://doi.org/10.1111/j.1745-3984.1992.tb00371.x>
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design* (ETS Research Report No. RR-03-16). ETS.
- Mohadjer, L., Berlin, M., Rieger, S., Waksberg, J., Rock, D., Yamamoto, K., Kirsch, I., & Kolstad, A. (1997). The role of incentives in literacy survey research. In A. C. Tuijnman, I. S. Kirsch, & D. A. Wagner (Eds.), *Adult basic skills: Innovations in measurement and policy analysis* (pp. 209-244). Hampton Press.
- Mohadjer, L., & Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, 26(2), 263-277. <https://doi.org/10.1108/QAE-06-2017-0031>

- Mosenthal, P. B., & Kirsch, I. S. (1991). Toward an explanatory model of document literacy. *Discourse Processes*, 14(2), 147–180. <https://doi.org/10.1080/01638539109544780>
- Mullis, I. V. S., & Martin, M. O. (2019). *PIRLS 2021 assessment frameworks*. TIMSS & PIRLS International Study Center, Boston College. <http://pirls2021.org/frameworks/>
- Mullis, I. V. S., Martin, M. O., & von Davier, M. (2021). *TIMSS 2023 assessment frameworks*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2023/frameworks/pdf/T23_Frameworks.pdf
- National Center for Education and the Economy. (2020, March 19). *Groundbreaking Maryland education reform bill passes in midst of Coronavirus crisis*. <https://ncee.org/tucker-writing/groundbreaking-maryland-education-reform-bill-passes-in-midst-of-coronavirus-crisis/>
- Ng, H. L., Poon, C. L., & Pang, E. (2020). Using IEA studies to inform policymaking and program development: The case of Singapore. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 245–260). Springer Nature. https://link.springer.com/chapter/10.1007/978-3-030-53081-5_14
- O'Dwyer, L. M., & Paolucci, C. (2019). Challenges in practice: A critical examination of efforts to link teacher practices and student achievement. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 471–491). SAGE. <https://doi.org/10.4135/9781526470379.n26>
- Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international large-scale assessment claims and country aims* (ETS Research Report No. RR-18-27). ETS. <https://doi.org/10.1002/ets2.12214>
- Oranje, A., & Ye, L. (2014). Population model size, bias, and variance in educational survey assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 203–228). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-14/population-model-size-bias-variance-educational-survey-assessments-andreas-oranje-lei-ye?context=ubx&refId=8bfdb5cf-e50a-41d3-9738-7e583589a86d>
- Organisation for Economic Co-operation and Development. (1992). *Adult illiteracy and economic performance*.
- Organisation for Economic Co-operation and Development. (2016). *Survey of adult skills technical report* (2nd ed.). https://www.oecd.org/skills/piaac/PIAAC_Technical_Report_2nd_Edition_Full_Report.pdf
- Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA2015_TechRep_Final.pdf
- Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results (Volume II): Where all students can succeed*. <https://doi.org/10.1787/b5fd1b8f-en>
- Organisation for Economic Co-operation and Development. (2020). *PISA 2018 technical report*. <https://www.oecd.org/pisa/data/pisa2018technicalreport/>

- Pepper, D. (2020). When assessment validation neglects any strand of validity evidence: An instructive example from PISA. *Educational Measurement: Issues and Practice*, 39(4), 8-20. <https://doi.org/10.1111/emip.12380>
- Pokropek, A. (2016a). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, 41(3), 300-325. <https://doi.org/10.3102/1076998616636618>
- Pokropek, A. (2016b). Introduction to instrumental variables and their application to large-scale assessment data. *Large-Scale Assessments in Education*, 4, Article 4. <https://doi.org/10.1186/s40536-016-0018-2>
- Pons, X. (2017). Fifteen years of research on PISA effects on education governance: A critical review. *European Journal of Education*, 52(2), 131-144. <https://doi.org/10.1111/ejed.12213>
- Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education*, 9, Article 10. <https://doi.org/10.1186/s40536-021-00104-6>
- Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, 9, Article 1. <https://doi.org/10.1186/s40536-020-00092-z>
- Ritzen, J. (2013). International large-scale assessments as change agents. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 13-24). Springer. https://doi.org/10.1007/978-94-007-4629-9_2
- Robinson, J. P. (2014). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 521-546). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-28/causal-inference-comparative-analysis-large-scale-assessment-data-joseph-robinson?context=ubx&refId=eb755322-c67a-4e25-9237-854b5df556d1>
- Rosdahl, A. (2014). *Fra 15 til 27 år: PISA 2000-eleverne I 2011/12 [From 15 to 27 years old: The PISA 2000 students in 2011/12]*. SFI, Det Nationale Forskningscenter for Velfærd. <https://www.vive.dk/media/pure/5070/276298>
- Rowley, K. J., Edmunds, C. C., Dufur, M. J., Jarvis, J. A., & Silveira, F. (2020). Contextualising the achievement gap: Assessing educational achievement, inequality, and disadvantage in high-income Countries. *Comparative Education*, 56(4), 459-483. <https://doi.org/10.1080/03050068.2020.1769928>
- Rust, K. (2014). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 117-154). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-11/sampling-weighting-variance-estimation-international-large-scale-assessments-keith-rust?context=ubx&refId=6b706f02-cc7d-4143-abaf-40619d1a0448>

- Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education*, 4, Article 6. <https://doi.org/10.1186/s40536-016-0019-1>
- Rutkowski, D., Thompson, G., & Rutkowski, L. (2020). Understanding the policy influence of international large-scale assessments in education. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (pp. 261–277). Springer Nature. https://doi.org/10.1007/978-3-030-53081-5_15
- Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). Assessment design for international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 75–95). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-9/assessment-design-international-large-scale-assessments-leslie-rutkowski-eugene-gonzalez-matthias-von-davier-yan-zhou?context=ubx&refId=0ea1bb81-169f-4ce1-a73f-b05180363b3c>
- Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, 45(4), 252–257. <https://doi.org/10.3102/0013189X16649961>
- Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research*, 62(3), 354–367. <https://doi.org/10.1080/00313831.2016.1261044>
- Rutkowski, L., & Rutkowski, D. (2019). Methodological challenges to measuring heterogeneous populations internationally. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 126–140). SAGE. <https://doi.org/10.4135/9781526470379.n8>
- Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators & fit measure performance. *Applied Measurement in Education*, 30(1), 39–51. <https://doi.org/10.1080/08957347.2016.1243540>
- Rutkowski, L., & Zhou, Y. (2015). The impact of missing and error-prone auxiliary information on sparse-matrix sub-population parameter estimates. *Methodology*, 11(3), 89–99. <https://doi.org/10.1027/1614-2241/a000095>
- Rychen, D. S., & Salganik, L. H. (Eds.). (2001). *Defining and selecting key competencies*. Hogrefe & Huber.
- Rychen, D. S., & Salganik, L. H. (2003). *Key competencies for a successful life and a well-functioning society*. Hogrefe & Huber.
- Sands, A., Goodman, M., Kirsch, I., & Dreier, K. (2021). *Opportunity across the states*. ETS. <https://www.ets.org/s/research/pdf/opportunity-across-the-states.pdf>
- Schleicher, A. (2018). *World class: How to build a 21st century school system*. OECD. https://www.oecd-ilibrary.org/world-class_5j8v15v201hd.pdf?itemId=%2Fcontent%2Fpublication%2F9789264300002-en&mimeType=pdf
- Schwippert, K. (Ed.). (2007). *Progress in reading literacy: The impact of PIRLS 2001 in 13 countries*. Waxmann.

- Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy*, 28(5), 710–725. <https://doi.org/10.1080/02680939.2013.779791>
- Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science*, 360(6384), 38–40. <https://doi.org/10.1126/science.aar4952>
- Singer, J. D., Braun, H. I., & Chudowsky, N. (Eds.). 2018. *International education assessments: Cautions, conundrums, and common sense*. National Academy of Education. <https://naeducation.org/wp-content/uploads/2018/08/International-Education-Assessments-NAEd-report.pdf>
- Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education*, 9, Article 8. <https://doi.org/10.1186/s40536-021-00100-w>
- Strietholt, R., & Rosén, M. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives*, 14(1), 1–26. <https://doi.org/10.1080/15366367.2015.1112711>
- Strietholt, R., & Scherer, R. (2018). The contribution of international large-scale assessments to educational research: Combining individual and institutional data sources. *Scandinavian Journal of Educational Research*, 62(3), 368–385. <https://doi.org/10.1080/00313831.2016.1258729>
- Suter, L. E. (2019). Growth and development of large-scale international comparative studies and their influence on comparative education thinking. In L.E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 197–223). SAGE. <https://doi.org/10.4135/9781526470379.n12>
- Suter, L. E., Smith, E., & Denman, B. D. (Eds.). (2019). *The SAGE handbook of comparative education studies*. SAGE Publishing. <https://doi.org/10.4135/9781526470379>
- Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, 67(1), 33–48. <https://link.springer.com/article/10.1007/BF02294708>
- Thurlow, M. L. (2014). Instructional and assessment accommodations in the 21st century. In L. Florian (Ed.), *The SAGE handbook of special education* (Vol. 2, pp. 597–631). SAGE Publications. <https://doi.org/10.4135/9781446282236.n37>
- Tucker, M. (Ed.). (2011). *Surpassing Shanghai: An agenda for American education built on the world's leading systems*. Harvard Education Press.
- Ulitzsch, E., von Davier, M., & Pohl, S. (2019). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, 55(3), 425–453. <http://doi.org/10.1080/00273171.2019.1643699>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- van de Vijver, F. J. R. (2018). Towards an integrated framework of bias in noncognitive assessment in large-scale international studies: Challenges and prospects. *Educational Measurement: Issues and Practice*, 37(4), 49–56. <https://doi.org/10.1111/emip.12227>

- van de Vijver, F. J. R., Jude, N., & Kuger, S. (2019). Challenges in large-scale education surveys. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative education studies* (pp. 83–102). SAGE Publications. <https://doi.org/10.4135/9781526470379.n6>
- Vera-Toscano, E., Rodrigues, M., & Costa, P. (2017). Beyond educational attainment: The importance of skills and lifelong learning for social outcomes. Evidence for Europe from PIAAC. *European Journal of Education*, 52(2), 217–231. <https://doi.org/10.1111/ejed.12211>
- Volante, L. (2013). Canadian policy responses to international comparison testing. *Interchange*, 44(3-4), 169–178. <https://doi.org/10.1007/s10780-014-9205-7>
- Volante, L., Fabio, X., & Ritzen, J. (2017). The OECD and educational policy reform: International surveys, governance, and policy evidence. *Canadian Journal of Educational Administration and Policy*, 184, 34–48.
- von Davier, M. (2014). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Chapman & Hall/CRC Press. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-13/imputing-prociency-data-planned-missingness-population-models-matthias-von-davier?context=ubx&refId=f5f1b7e1-533e-40e1-95e9-54dc43ef3848>
- von Davier, M., & Bezirhan, U. (2022). A robust method for detecting item misfit in large-scale assessments. *Educational and Psychological Measurement*, 83(4), 740–765. <https://doi.org/10.1177/00131644221105819>
- von Davier, M., Gonzalez, E., & Schulz, W. (2020). Ensuring validity in international comparisons using state-of-the-art psychometric methodologies. In H. Wagemaker (Ed.), *IEA research for education: Vol. 10. Reliability and validity of international large-scale assessment* (pp. 187–220). Springer. https://doi.org/10.1007/978-3-030-53081-5_11
- von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, 44(6), 671–705. <https://doi.org/10.3102/1076998619881789>
- von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*, 42(4), 291–306. <https://doi.org/10.1177/0146621617730389>
- von Davier, M., & Sinharay, S. (2014). Analytics in international large scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Chapman & Hall/CRC Press. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-12/analytics-international-large-scale-assessments-item-response-theory-population-models-matthias-von-davier-sinharay?context=ubx&refId=d996833d-7903-4b7a-ac0b-d0799cb098cd>

- von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, 26(4), 466–488. <https://doi.org/10.1080/0969594X.2019.1586642>
- Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–36). Chapman & Hall/CRC. <https://www.taylorfrancis.com/chapters/edit/10.1201/b16061-6/international-large-scale-assessments-research-policy-hans-wagemaker?context=ubx&refId=51ae00d5-8ff4-4252-9d46-ab7232d19820>
- Wetzel, E., Xu, X., & von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement*, 75(5), 739–763. <https://doi.org/10.1177/0013164414558843>
- Wirtz, W., & Lapointe, A. (1982). Measuring the quality of education: A report on assessing educational progress. *Educational Measurement: Issues and Practice*, 1, 17–19, 23. <https://doi.org/10.1111/j.1745-3992.1982.tb00673.x>
- Wise, S. L. (2020). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Applications of artificial intelligence (AI) to assessment* (pp. 161–174). Information Age Publishing.
- Wiseman, A. (2013). Policy responses to PISA in comparative perspective. In H. D. Meyer & A. Benavot (Eds.), *PISA, power, and policy the emergence of global educational governance* (pp. 303–322). Symposium Books.
- Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, 26(2), 196–212. <https://doi.org/10.1108/QAE-07-2017-0038>
- Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, 37(4), 16–27. <https://doi.org/10.1111/emip.12226>
- Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer. <https://doi.org/10.1007/978-3-319-56129-5>

NOTES

1. In this chapter, we use the term *country* to signify a participating jurisdiction for which results are reported. This term includes “countries” as commonly understood but may also refer to other types of political entities. In ILSA reports, the latter are variously described as *economies* or *benchmarking participants*.
2. Human capital is often characterized as a broad set of cognitive and noncognitive skills and knowledge that is necessary in modern economies. See Kirsch et al. (2016) for a discussion of the growing importance of human and social capital and their connections to opportunity.

3. Advanced TIMSS was intended to assess students close to the end of compulsory education (e.g. Grade 12 in the United States). It was administered in 1995, 2008, and 2015.
4. Each paying participant obtains a database with its results, and the results are published in the accompanying reports.
5. More exactly, PIRLS is administered to students in the upper of the two grades that enroll the most 9-year-olds. This is the fourth grade in most countries.
6. For PIRLS 2021, ePIRLS booklets were more fully integrated into the booklet structure of digital PIRLS.
7. These assessments were the Young Adult Literacy Survey 1984, the Department of Labor Study of JTPA Trainees and UI recipients 1992, and the National Adult Literacy Survey 1993.
8. The ICILS was conducted in 2013, 2018, and 2023.
9. ICCS assesses students enrolled in the eighth grade, provided that the average age of students at this year level is 13.5 years or above. In countries where the average age of students in Grade 8 is less than 13.5 years, Grade 9 is defined as the target population. The ICCS was conducted in 2009, 2016, and 2022.
10. Beginning in 2003, succeeding assessments retained the acronym TIMSS, with “Trends” replacing “Third” in the title.
11. BICSE was established in 1988 and disbanded in 2002. Throughout its existence, the National Center for Education Statistics provided substantial funding.
12. With a statutory mandate to collect data in the United States, the National Center for Education Statistics, a division within the U.S. Department of Education, has also provided crucial, sustained support for many of the international assessments beyond the contribution required for U.S. participation.
13. These dates mark the beginning of data collection for the surveys. Of course, preparatory work began years earlier.
14. For a contemporary treatment of constructs and test scores, see Haertel (2018).
15. This and other issues are reviewed as part of the adjudication process that is completed at the end of the project cycle.
16. The separate estimation of item parameters for a country (or set of countries) improves the estimated score distributions for those countries. If the number of such items is small, this policy should not materially impact the comparability of the described proficiency scale across all countries. However, if the number of items requiring separate estimation exceeds a certain threshold (yet to be determined), then it may be advisable to construct a unique scale for that country (based on its data only) and sacrifice comparability with the main scale.
17. For a more complete discussion of this activity, see Dept et al. (2025).
18. For example, in school-based surveys, the aim is to estimate score distributions for students, but the school is a relevant level of analysis. Consequently, the

sample design must allow for selecting multiple students within each school, as well as large numbers of schools. Schools constitute the first stage of the design, and students within schools constitute the second stage.

19. To prepare for school nonresponse, replacement schools are selected at the same time as the main sample.
20. For example, in PIRLS 2016 the target population was students in the fourth year of formal schooling. In most countries this was Grade 4. However, if the mean age in Grade 4 was less than 9.5 years, Grade 5 was chosen. For further details, see Martin et al. (2017, chap. 3).
21. If the school has fewer than 35 eligible students, all eligible students are selected for administration.
22. As noted earlier, countries vary in the proportions of the birth cohorts enrolled in school, with implications for the kinds of (comparative) inferences that can be made from the data collected.
23. The problem has become more salient with the greater participation of middle- and low-income countries.
24. Blocks are systematically varied by position within booklets to remove any order effects.
25. More exactly, the students responding to each item have known probabilities of selection so that the raw results can be modified by appropriate weighting. With adaptive testing, the students exposed to different item blocks are not randomly equivalent. Nonetheless, the designs do enable unbiased estimation of item parameters.
26. In many contexts, such item sets are called *testlets*.
27. Items requiring human scoring cannot contribute to the real-time decision-making process; however, they may be essential for construct representation.
28. When there is interest in reporting information at the subscale level, items are grouped by subscale and the item parameters are obtained from the unidimensional scaling employed in scale construction. This strategy possesses advantages when comparing subscale profiles across countries.
29. For more technical detail on addressing item–country interactions, see, for example, Martin et al. (2016, chap. 11) for TIMSS 2015, and Martin et al. (2017, chap. 10) for PIRLS 2016. On occasion, a single separate item calibration is carried out for a group of countries.
30. See also Mullis et al. (2021, chap. 4).
31. As a general rule, the number of principal components retained is limited to no more than 5% of a country's student sample size, thereby reducing the percentage of variance accounted for to avoid overspecification of the conditioning model. (See Martin et al., 2020, chap. 12.)
32. Various versions of the expectation–maximization algorithm are used to obtain parameter estimates for the latent regression model (von Davier & Sinharay, 2014).

33. This discussion assumes there are no missing data in the background questionnaire, an assumption that is rarely fulfilled. Recent research has investigated methods for carrying out the estimation procedures in the face of missing data (e.g., Grund et al., 2021).
34. See also OECD, 2017, chap. 9.
35. More exactly, the key requirement is that there is sufficient information (i.e., large enough sample size) on each trend item so that it is possible to judge whether it is working well in each population for which results are reported—typically a (large) language group within a country.
36. In this context, “freeing item parameters” entails estimating the item parameters separately for each country, rather than having one set of item parameters for all countries. For more technical detail on addressing item-by-country interactions, see Martin et al. (2020, chap. 10) for TIMSS 2019 and Martin et al. (2017, chap. 10) for PIRLS 2016.
37. See also OECD (2017, chap. 9).
38. NAEP no longer uses this approach, having shifted to employing achievement-level descriptors.
39. More generally, in PISA, students who are administered items in the major domain and only one of the minor domains receive PVs for the other minor domain.
40. Recall that sampling weights are used to control the proportional representation of the cases in the estimation of population parameters.
41. See International Association for the Evaluation of Educational Achievement (n.d.-b).
42. Both tools are considered proprietary software developed specifically for use with ILSA data. The current version of the microdata analyzer is called IDBA.
43. Attendees may be required to pay fees to attend, depending on the level of governmental support available.
44. See International Association for the Evaluation of Educational Achievement (n.d.-a).i
45. See, for example, Martin et al. (1999).
46. Canada, France, Norway, Switzerland, England, and Shanghai.
47. Educational attainment is typically quantified in terms of either years of schooling completed or reaching certain milestones, such as completion of secondary school.
48. Developing an appropriate cross-national measure of socioeconomic status is itself a challenging endeavor (Avvisati, 2020; Marks & O’Connell, 2021).
49. One concern with such studies is that the focal construct may be operationally defined differently in each assessment. Such a “construct shift” limits the kinds of conclusions that can be drawn.
50. Although the transition from paper-and-pencil administration to digitally based assessments occurs at a point in time, it may prove necessary to replenish the

paper-and-pencil administration item pools if there is an ongoing need to offer the paper-and-pencil administration option.

51. Maintaining item parameter invariance is one way to achieve such comparability.
52. The lack of familiarity is mitigated somewhat with tutorials and practice exercises that familiarize respondents with new item response formats.
53. The issue is complicated by the fact that the construct definitions and the corresponding assessment frameworks do not always make explicit which types of prior knowledge are part of the construct and which are not.
54. Denmark, Finland, Iceland, Norway, and Sweden.
55. In the ILSA context, an extreme case of the aptitude–achievement paradox occurs when the relationship between a cognitive outcome and a behavioral (or attitudinal) outcome is positive within countries but negative among countries. See Ainley and Ainley (2019) for further discussion.
56. Some countries employed incentives in Cycle 1 of PIAAC and others are planning to use them in Cycle 2. An incentive study conducted in conjunction with the U.S. National Adult Literacy Study found that monetary incentives increased participation rates but not performance (Mohadjer et al., 1997).
57. Cognitive laboratories were conducted for PIAAC, with particular attention on interface issues related to usability. However, they are not standard for school-based ILSAs.
58. We thank Xue Jiang and Shinji Katsumoto, Teachers College, Columbia University, for the translations.
59. To this point, M. Smith (cited in Singer et al., 2018, p. 28, footnote) asserted that in the past, legislators found state-level differences more compelling than country-level differences.
60. Quoted in Singer et al. (2018), pp. 28–29.
61. Consult Maehler et al. (2020) for a comprehensive bibliography.
62. For an extended example using U.S. PIAAC data, see the report by Sands et al. (2021).
63. For more on Shanghai’s PISA results, see Loveless (2014).

Part V

Measurement as a Science

On the Nature of Measurement

Derek C. Briggs

University of Colorado Boulder

Andrew Maul

University of California, Santa Barbara

Joshua A. McGrane

Assessment & Evaluation Research Centre,
The University of Melbourne

Measurement, in its many forms, has been an integral part of human civilization since its very inception and is now inextricably woven into nearly every aspect of our lives. Every time we eat food, drive a car, take medicine, check a weather report, make a financial transaction, or engage with society in nearly any other way, we place our trust in the results of measurements. And for the most part, that trust seems to be well earned; over centuries of development, particularly in the contexts of the physical sciences and engineering, the concept of measurement has come to be commonly associated with precision, accuracy, objectivity, and overall trustworthiness.

Against this backdrop, it seems little wonder that the language and logic of measurement has been applied in nonphysical contexts as well. Discourse on measurement has featured prominently in the social and human sciences since their formal inception in the late 19th century, and arguably in economics for even longer than that. Moreover, the language and logic of measurement have often been employed, with varying extents of formality, in relation to *testing*, especially in vocational and educational contexts. This is visible in many ways but is perhaps most obvious when the results of tests are reported in quantitative terms and when the differences between test scores (between or within students) are interpreted in terms of magnitudes, implying meaningfulness of the units of the scales on which the results are reported. We suspect most readers are familiar with many examples of educational test results being interpreted as *measures* of students' knowledge, skills, and abilities (and perhaps, by extension, as measures of aspects of teacher and school quality, or the efficacy of educational programs and interventions), as well as results from surveys and other modes of assessment being interpreted as measures of a range of psychological and social properties (e.g., conscientiousness, self-control, empathy).

However, as we will discuss in this chapter, testing is not necessarily the same as measurement, and the mere existence of a systematic number-generating procedure is insufficient to ground subsequent claims about numeric values to the kinds of epistemic virtues—that is, precision, accuracy, objectivity, and resulting trustworthiness—commonly associated with measurement. This distinction is often not clearly articulated, including, arguably, in previous editions of this volume. As was pointed out by Borsboom (2009, p. 706) in a review of the fourth edition of *Educational Measurement*,

the authors . . . appear to consistently conflate the terms *testing* and *measurement* . . . as if any test score is automatically a measure. . . . Test scores can sometimes be interpreted as measurements . . . but not always and not necessarily so. The question of whether educational test scores can be so interpreted, and under which conditions, receives no attention at all in *Educational Measurement*—which must be considered remarkable.

Stated as simply as possible, a key purpose of the present chapter is to engage with the question of how educational measurement can (and perhaps should) be understood with respect to how measurement is understood more broadly, given the weight this concept carries in terms of interpretations of objectivity and trustworthiness. In doing

so, we do not intend to argue for a single set of necessary and sufficient conditions for an activity to be considered an instance of measurement. We will, however, argue that there are common components of trustworthy measurement processes across domains of application and that recognizing these commonalities can help frame and clarify the central issue of how measurement results are interpreted and how their claimed trustworthiness is justified.

Broadly, and in accordance with the metrological literature (e.g., Mari et al., 2017, 2023), we define measurement as *a process that provides information about a property of an object or event in the world, uses instrumentation designed to be sensitive to differences in this property, produces information in the form of values related to a scale or unit for that same property, and is qualified by information about uncertainty*. In our view, the elements of this process apply regardless of whether the measured property¹ is exhaustively definable in physical terms (e.g., temperature) or not (e.g., reading comprehension ability); our contention is that whether one is speaking of measurement in an educational context, in psychology, or in the physical sciences, the activity will involve a common set of elements that serve to justify its claimed trustworthiness. By drawing attention to these common elements, the challenges and affordances of measurement in a specific domain of application become more apparent.

It is worth noting that there is substantial variation in definitions of measurement throughout the historical and contemporary literature, and here we note just a few. Perhaps most famously, Stevens (1946, p. 667) defined measurement as “the assignment of numerals to objects or events according to rules.” Michell (e.g., 1999, p. 222) defined measurement as the “discovery or estimation of the ratio of a magnitude of a quantity to a unit of the same quantity.” Writing in the context of educational testing, Lord and Novick (1968, p. 17) defined measurement as “a procedure for the assignment of numbers (scores, measurements) to specified properties of experimental units in such a way as to characterize and preserve specified relationships in the behavioral domain.” According to Mislevy (2018, p. 313), “Educational measurement, conceived broadly, is situating observations from an assessment into a mathematical framework to support reasoning about peoples’ capabilities.” Finally, a 2022 task force commissioned for the National Council on Measurement in Education (NCME) produced a report with two distinct definitions for measurement and educational measurement. According to the task force,

Measurement is a systematic process of data collection using instrumentation that results in a quantity intended to support inferences about an attribute or property of an object, event, or phenomenon. Educational measurement involves measurement of knowledge, skills, dispositions, and abilities for some educational purpose, such as supporting learning, certifying learning, or identifying policies and practices that improve learning. (Ackerman et al., 2024, p. 8)

Inspection of and debate over what has been included and excluded in any given definition of measurement can lead to lively exchanges, whether in a classroom setting or in

a pub. It may not always be apparent that different definitions can imply very different philosophical commitments about measurement as a process (c.f., Michell, 1999; Stevens, 1946), and this is something we explore in the “Historical and Conceptual Background” section of this chapter. At the same time, any one-sentence definition of measurement can only be a starting point for a more complete understanding. In our view, whenever we call something “measurement,” we are at least implicitly laying claim to the perceived objectivity and trustworthiness commonly associated with this term. The shorthand definition that we favor is intended to draw attention to the components of a measurement process that serve to underwrite this claim. To go deeper, in this chapter we propose a framework for understanding measurement with respect to four central components: theory (T), instrumentation (I), scales and units (S), and modeling (M). We refer to this as the TISM framework, and we argue that, regardless of the subject matter or domain of application, the practice of measurement requires attention to each of these four components and their interactions.

The balance of this chapter is organized as follows. In the section “Historical and Conceptual Background,” we present an overview of some of the key historical developments that have taken human societies from a time when the terminology of measurement was primarily reserved for applications in the physical sciences to the early 21st century and the context of this edited volume, which is concerned with the methods and practices of educational measurement. We explain how this more broadened conceptualization of measurement came about. We also summarize some of the hallmarks of educational measurement: its practical focus on decision-making, its association with standardized testing, and its emphasis on treating the results of a test as a random variable amenable to statistical modeling.

In the section “Measurement and the TISM Framework,” we present and illustrate the central contribution of this chapter, the TISM framework for measurement. We describe each component of the framework and then illustrate the components in two concrete measurement contexts. We do this first for the canonical example of measuring temperature. We then shift to a more detailed illustration in an educational context for the measurement of reading comprehension. We show that the four TISM components of measurement are present and relevant in both contexts, but tend to pose unique challenges in education. We argue that this is because our theoretical understandings of constructs of measurement are often nascent or in flux and, relatedly, because the scales and units we use to interpret measurement results often have a tenuous connection to this theoretical understanding.

In the section “Challenges and Debates in Educational Measurement,” we discuss how the TISM framework relates to validity and validation. We first briefly review the literatures on these topics and discuss how a focus on measurement (and the TISM components in particular) can be viewed as compatible with and complementary to an argument-based approach to validation: Stated briefly, whenever an argument for the adequacy of a test for a given purpose involves a claim about measurement, the TISM framework helps clarify what kinds of issues would be necessary to attend to

in order to evaluate and defend this argument. We then illustrate this by turning to two examples of educational measurement activities that are challenging because, in our view, they depend on the extent to which one has an adequate substantive theory of the construct (the T in our TISM framework), whether this theoretical conceptualization coheres with the construction of test instruments (I), the scale and unit on which test results are expressed (S), and the choice of a statistical model used to relate observed test scores to the construct of interest (M). Specifically, we consider the activities involved in (a) characterizing the reliability, measurement error, and uncertainty of a measuring instrument and (b) supporting inferences about magnitude relative to the units of a measuring scale. We conclude the chapter with a summary of our main points, an acknowledgment of an important limitation of the TISM framework, and some speculation about the future of educational measurement.²

HISTORICAL AND CONCEPTUAL BACKGROUND

Measurement in the Physical Sciences

Measurement has a long history, stretching back at least 2,500 years across a multitude of cultures. Inevitably, thinking about measurement and its role in science and society has evolved concomitantly with science and society in general. A first observation about the history of measurement is that the vast majority of this history is that of physical measurement; for the most part, it has only been since the late 19th century that measurement-related concepts have been formally incorporated into the study of social, psychological, and educational phenomena. For this reason, much of the vocabulary of educational measurement has been borrowed from the physical sciences, and more specifically, from physics and engineering (McGrane & Maul, 2020). Within these disciplines, a dominant (though not universally accepted) conception of measurement draws from Euclid's *Elements*, sometimes thought of as “the earliest contribution to the philosophy of measurement available in the historical record” (Michell, 2005, p. 288). Euclid defined *measure* in terms of ratios of magnitudes: “A magnitude is a part of a[nother] magnitude, the lesser of the greater, when it *measures* the greater” (emphasis added; Euclid, trans. 2008, p. 130). The process of *measurement* then became understood as the estimation of ratios of magnitudes of empirical properties (see, e.g., Michell, 1999).

The impact of this “Euclidean” understanding of measurement is still evident in the writings of physical scientists more than 2 millennia after Euclid. A paradigmatic example is that of the Scottish physicist James Clerk Maxwell, who opened his 1873 book *Treatise on Electricity and Magnetism* with a section subtitled “On the Measurement of Quantities”:

Every expression of a Quantity consists of two factors or components. One of these is the name of a certain known quantity of the same kind as the quantity to be expressed, which is taken as the standard reference. The other component is the number of times the standard is to be taken in order to make up the required quantity. The standard quantity is technically called the Unit, and the number is called the Numerical Value of the quantity. (Maxwell, 1873, p. 1)

Central to this understanding of measurement is the conviction that magnitudes of quantities exist and that measurement is thus a process of discovery of mind-independent facts about their ratios; this perspective is consistent with *scientific realism*, which, broadly, asserts that there is an objective, spatiotemporally structured world and that at least one of the aims of scientific inquiry is the acquisition of knowledge about this world (see, e.g., Michell, 2005; Rosenberg & McIntyre, 2020). These realist views were (and still are) integral to the development of physics as a quantitative science; to a large extent, the advancement of physics has entailed the discovery of networks of lawful relationships among physical quantities, taken to be invariant features of the physical world.

Maxwell's work laid the foundation for the International System of Quantities, which consist of seven "base" or "fundamental" quantities (so called because they cannot be expressed as a function of other fundamental quantities: length, mass, duration, intensity of electric current, thermodynamic temperature, amount of substance, and luminous intensity) from which other physical quantities may be derived through physical laws (e.g., velocity as the ratio of length and time). Maxwell also helped pioneer the movement to establish the meaning of units of fundamental quantities in terms of invariant physical phenomena—for example, the meter, once defined by the length of a specific physical rod, is now defined in terms of the distance light travels in a vacuum in a given time duration—which facilitates efforts to make measurement results *subject independent*, that is, interpretable in the same way by everyone, regardless of place or time (see, e.g., Maul et al., 2019; McGrane, 2015).

Measurement in the Human Sciences

Some of the earliest applications of measurement to nonphysical human properties were in psychophysics, a program of research that originated through the contributions of Ernst Weber and Gustav Fechner in the mid- to late 19th century (see Briggs, 2021b, for an extensive overview of this history). The psychophysicists were chiefly concerned with the measurement of different kinds of "sensory sensation" through the empirical establishment of law-like relationships (e.g., Fechner's law) between the quantitative levels of a physical stimulus and levels of psychical sensation. Two distinguishing features of the psychophysics tradition came in the use of controlled experiments to provide empirical evidence for these "laws" and in the use of probability theory in general, and the normal cumulative distribution function in particular, to relate sensory "errors" to levels on a sensation continuum. Both features were inspirational in the early development of what came to be known first as latent trait theory and then eventually item response theory (G. A. Ferguson, 1942; Guilford, 1936; Lawley, 1943; Lord, 1952; Mosier, 1940, 1941; Tucker, 1946).

Louis Thurstone reconceptualized Fechner's law as a special case of his "law of comparative judgment" (Thurstone, 1927a, 1927b, 1927c). In contrast to Fechner and classical psychophysics, Thurstone's approach permitted qualitative stimuli as a starting point, so long as the stimuli could be ordered. Thurstone saw that the same mathematical psychophysical models could be applied to discriminating between the weight

of two objects (with weight being a purely physical property), or the legibility of two samples of handwriting (the perception of which involves human psychology). In the former case, the order of the stimuli was known in advance; in the latter case it would need to be inferred from the experiment. But in either case, if it could be established empirically that errors in judgment were well approximated by a normal distribution, a scale with units defined by standard deviations could be established. Although Thurstone showed little interest in metaphysical speculation about the nature of the psychological construct being measured via the law of comparative judgment, just as it had for Fechner, the approach hinged on the assumption that whatever the construct was, if it could be ordered by observational comparison, methods could be used to represent it on a continuum with a meaningful scale.

Another contemporaneous tradition that drove applications of measurement to nonphysical properties is now referred to as “individual differences” research, so called because of its shift of attention to between-person (i.e., *interindividual*) variability as a subject for study, whereas the psychophysicists were primarily concerned with within-person (i.e., *intraindividual*) variability (again, see Briggs, 2021b, for an overview of this history). Francis Galton was the chief architect of this approach in the late 19th century based on his assumption that what seemed to be true of human physical properties such as height and weight—that is, that they were normally distributed—would also be true of psychological constructs, whether conceptualized narrowly, as the ability to discriminate lengths (as in psychophysics), or more broadly, as the ability to solve mathematical problems in an educational context. Galton would justify this assumption as the consequence of a speculative theory about human heredity, positing that all human attributes are the result of a large number of small, independent random events that accumulate and thereby explain why individuals differ with respect to any particular physical or psychological characteristic.

Galton proposed a strategy for his research that he described as “relative” measurement, which only depended on the ability of an observer to accurately rank individuals with respect to some designated measurement construct. If the construct to which the ranks applied was assumed to follow a normal distribution, then the ranks could be converted into percentiles and transformed using the normal ogive into a location on a “statistical” scale with units defined by standard deviations (Galton, 1875). With this view, the primary challenge of measurement was an instrumental one. By 1884, Galton had introduced a proof of concept for how such an undertaking could be accomplished after opening an anthropometric laboratory at the site of London’s International Health Exhibition. For just 3 pence and 20 to 30 minutes of time, 9,337 visitors to the laboratory emerged with measurements of their physical attributes (length and weight), motor constructs (breathing and strength), and sensory constructs (visual acuity, color sense, discrimination of length, and discrimination of angles) recorded on a single sheet of paper. As Galton wrote, “One of the conditions that a standard set of instruments ought to fulfill is that it should admit to being readily packed, carried about

from place to place, and quickly set up anywhere for temporary use by a professional measurer” (Galton, 1885, p. 207).

Galton had advanced the position that mental tests of sensory discrimination were a promising way to measure at least some aspect of human intelligence. It was in the investigation of this position that Charles Spearman introduced many pivotal concepts and methods that remain at the heart of the study of individual differences and the practice of educational measurement (Spearman, 1904, 1907, 1910). In particular, Spearman formalized a distinction between the value of a human construct that one wishes to measure, the value that one observes as the outcome of a measurement procedure, and the difference between the two, attributable to chance error. This was the genesis of the linear error model that remains at the heart of classical test theory (for details, see Clauser et al., this volume; Briggs, 2021b; Clauser, 2021), and an elaborated version of it was at the heart of what Spearman regarded as a model for mental measurement, his theory of two factors.

Spearman (1904) had postulated that performance on any mental test could be modeled as a linear function of two quantitative mental attributes: one that was specific to that test and one that was general to all tests—the latter representing the famous “general intelligence factor,” or *g*. Spearman’s theory of two factors was intended to explain the interesting patterns he had observed among the correlations computed for teachers’ grades across a range of subjects and, later, across batteries of early mental tests. In this regard, Spearman’s theory was agnostic about the instrumentation used to elicit differences in intelligence. In fact, Spearman viewed this as an important feature of his approach to measurement, because it made no *a priori* assumptions about the quality of any given test as a measure of intelligence. Instead, the quality of a test could be judged “objectively” by ascertaining its loading on *g* relative to competing mental tests.

Taking a far more instrumental tact that would usher in a whole new wave of non-physical measurement in the 20th century by way of “intelligence testing,” Alfred Binet, in collaboration with Theodore Simon, developed clever “tests” that were to be administered as part of a short oral interview for the purpose of identifying French children with learning disabilities. In contrast to prevailing mental tests of the time (Cattell, 1890), which emphasized sensorimotor discrimination, Binet’s tests targeted a variety of complex cognitive abilities related not just to sensory discrimination but also to memory, comprehension of language, abstract reasoning, metacognition, and adaptation. Each test was meant to be engaging for the child and easy to administer, but also to provide actionable diagnostic insights into the multifaceted nature of a child’s developing intellect.

Binet referred to his procedure as a “measuring scale” for intelligence because by the time of the 1908 revision, one of its defining features was that it could be used to produce a numeric estimate of a child’s intellectual level expressed in years. This was accomplished in two stages: a first stage in which each test in Binet’s battery could be located with respect to the earliest age at which some proportion of children in a reference population could complete it successfully and a second stage where an individual child’s

intellectual level could be measured relative to this reference population as a cumulation of tests successfully completed. That is, if each test could be given an ordered location on a measuring scale, then the scale could be applied to individual students as if it represented an intellectual stadiometer (an instrument for measuring height). But this was meant to be largely metaphorical because Binet recognized that what he had constructed was a method for classification, and classification fell short of measurement because it did not involve the assessment of magnitude. (For details, see Briggs, 2021b.)

A Broadened Conceptualization of Measurement

In 1932, the British Association for the Advancement of Science appointed a committee (called the Ferguson Committee since it was chaired by the physicist A. Ferguson) to consider the scientific status of psychophysical measurement. The committee consisted of a mixture of psychologists and physicists, including the physicist Norman Campbell, who had recently published two influential books delineating a theory of measurement in the physical sciences (Campbell, 1920, 1928, 1938). Campbell argued that for *fundamental measurement* to be possible, it had to be possible to demonstrate that a property was both orderable and additive. By additive, Campbell meant that given two objects A and B with a common property, it must be possible to form a third object C such that the property of C is the combination of A and B. Only when this could be satisfied would it be defensible to develop a procedure for representing properties of objects with numbers.

Given this understanding of measurement, psychophysics could not be considered an instance of fundamental measurement because there was no way to show that “sensory sensation” could satisfy the requirement of additivity. The Ferguson Committee met on multiple occasions over an 8-year span and released two reports indicative of a stalemate (A. Ferguson et al, 1938, 1940). The physicists on the committee argued that psychophysics was not measurement; the psychologists on the committee balked at the consequences of Campbell’s restrictive conception of measurement but failed to suggest any compelling alternative.

It was this outcome from the Ferguson Committee that prompted the psychologist Stanley Smith Stevens (1946, 1951) to propose a broadened definition of measurement as the assignment of numerals to objects or events according to rules. Stevens’s definition of measurement had accepted from Campbell the premise that the activity of measurement was one of numeric representation, rejected the premise that only additive properties were measurable, and dropped altogether the formal distinction between an object and a property of an object. Stevens argued that additivity was just one of many “rules” that could be used in any experimental procedure that required humans to relate numerals to objects or events. Stevens and other psychophysicists invoked a variety of operational procedures in their experiments to elicit judgments about physical stimuli with respect to order, difference, and ratio.

From here, Stevens made distinctions regarding the “level” of measurement that could be attained as a function of different rules for numerical assignment through a taxonomy of nominal, ordinal, interval, and ratio scales. Stevens also introduced the

concept of *permissible statistics*, arguing that the types of statistical analyses that could be justified depended on the strengths of the scales of the relevant measures. For example, with measures on an interval or ratio scale, it would be meaningful, and therefore permissible, to make comparisons premised on the computation of means and standard deviations, but not for measures on an ordinal or nominal scale. (For more on Stevens, his taxonomy, and his approach to measurement, see Briggs, 2021b; Luce & Suppes, 2002; McGrane, 2015; Michell, 1986, 1999; Stevens, 1946, 1951.)

The distinction Stevens was making between measurement and scaling to a great extent represented a codification of the changing practices of different fields of quantitative psychology between 1860 and 1950. In the influential work of Fechner, Thurstone, Galton, Spearman, and Binet, among others, the terminology historically associated with physical measurement was being extended to any attempt to subject human characteristics to quantitative treatment. While there may have been differences in focus in the methods proposed for the measurement of nonphysical human constructs, it is noteworthy that a consensus remained that the ideal kind of measurement was that which produced measures on what Stevens was defining as either an interval or, preferably, a ratio scale. Moving forward, Stevens's broadened definition became highly influential—it was the definition of measurement invoked by Lorge (1951) in the first edition of *Educational Measurement*, and it has, arguably, remained a tacitly accepted definition in education and psychology to the present day (Michell, 1997).

Measurement in Education

In the early 20th century, the practice of *educational measurement* became synonymous with the development of group-administered and objectively scorable tests of students' knowledge, skills, and abilities that were thought to be acquired in or relevant to school settings. The key technology that made educational measurement possible on a large scale came directly from advances in the technology of testing prompted by group administration and multiple-choice item formats, innovations that emanated from the demonstration that intelligence testing could be used to facilitate the efficient placement of military personnel during World War I (e.g., Jones & Thissen, 2006; Yerkes, 1921). Francis Galton had first collected measurements for roughly 9,000 individuals at his anthropometric laboratory (Briggs, 2021b); just 7 years later, Joseph Mayer Rice had given some 33,000 students common tests of spelling (Kridel, 2000). But nothing came close to the accomplishment of the Army Alpha and Beta exams, administered cumulatively to 1.75 million recruits by the time World War I ended in 1918 (Kevles, 1968).

In the pioneering works that had introduced mental testing to a broader audience (Cattell, 1890; E. L. Thorndike, 1904), the products of mental tests had been referred to as measurements, in (at least) the loose sense that every test could be used to assign a number to a person where no number had existed before. The “educational” modifier seems to have been added to distinguish the measurement of student attainment (E. L. Thorndike, 1912, 1918) from the measurement of the broader suite of psychological

constructs that had become the province of tests of intelligence, personality, and attitudes. Hence, if domain-specific tests of student attainment produced educational measurements (plural), the process of creating the tests combined with analysis of results for comparative purposes fell under the general purview of educational measurement (singular).³

Even before the Binet–Simon scale had been translated into English (Goddard, 1910) and adapted for use with American children (Terman, 1911, 1916), a growing interest in educational measurement had already begun to take root in the United States as part of a broader progressive movement. A key idea being advanced was that public education could be improved systematically by applying the successful methods of the physical sciences—in particular, measurement and experimentation—to the evaluation of curricular and instructional efficacy. Such an approach hinged on the availability of standardized tests to provide educational measurements of the desired “products” of education (E. L. Thorndike, 1912). This was at the heart of the quantitative imperative articulated most famously by E. L. Thorndike, the father of educational measurement, in his “credo”:

Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality. Education is concerned with changes in human beings; a change is a difference between two conditions; each of these conditions is known to us only by the products produced by it—things made, words spoken, acts performed, and the like. . . . To measure a product well means to define its amount [so] that competent persons will know how large it is, with some precision, and that this knowledge may be conveniently recorded and used. This is the general Credo of those who, in the last decade, have been busy trying to extend and improve measurements of educational products. (E. L. Thorndike, 1918, p. 16)

It might be easy to miss some important subtleties regarding E. L. Thorndike’s approach to educational measurement. To begin with, the “whatever” Thorndike took to exist as a quantity in educational contexts was not some educational or psychological attribute of a student, but the attribute of an educational product, such as the legibility of a sample of handwriting, the merit of a written essay, or the complexity of an arithmetic problem a student has been asked to solve. What could be known about student learning must be inferred from what could be observed in the products created by students and the relationship between the quality of these products and outcomes of societal value. The role of measurement in this context was to establish scales for the quality of these different products, so that the scales could be used to clarify and standardize goals for instruction and learning. In this sense, what Thorndike referred to as measurement might be what others would today refer to as “scaling.”

Educational measurement has always had a decidedly practical orientation because it is typically undertaken in support of evaluative decision-making with important ramifications for children and young adults in particular. For example, Rice (1913) was one of the first to use standard tests to demonstrate that students exposed to instruction with heavy doses of repetitive spelling and arithmetic drills were no more likely to excel

in their subsequent achievement than students with a lower dosage of this instruction. E. L. Thorndike had developed measuring scales for the quality of written composition to support the argument that drilling students in Latin grammar did not produce transferable skills that helped them become better writers (Woodworth, 1952). These were examples in which educational measurement was being used for comparative purposes in the aggregate. By the mid-20th century, however, a national testing industry had emerged with a focus on the use of educational measurements to also classify and sort *individual* students.

This practical orientation was evident in the chapters of the first edition of *Educational Measurement* (Lindquist, 1951) and has been evident in all the editions that have followed. As E. F. Lindquist wrote in this book's preface, "There has long been an urgent need, particularly in the advanced training of measurement workers, for a comprehensive handbook and textbook on the theory and technique of educational measurement" (Lindquist, 1951, p. vii). What Lindquist seems to have meant by "technique" were the nine chapters in Part 2 of the book devoted to "The Construction of Achievement Tests." This corresponds to the component of measurement that we have identified as instrumentation—the I in our TISM framework that we elaborate in the section of this chapter entitled "Measurement and the TISM Framework." However, what Lindquist meant by theory was not quite the same as the component we identify as the T in our TISM framework and describe as the need to have a substantive theory about the construct of measurement. Instead, the theory he had in mind was the developing wisdom on *test* theory—that is, the use of statistical models and methods to inform the optimal design of tests to achieve their intended purposes in educational contexts. It is to developments in mathematical and statistical modeling of test data that we turn to next in more detail because this represents an area in which the scholarship related to educational measurement has been especially active and productive since the mid-20th century.

Statistical Models and Educational Measurement

Classical Test Theory

Although the core concepts at the heart of classical test theory (CTT) can be said to have originated in the work of Charles Spearman and Francis Edgeworth near the turn of the 20th century, further refinement and formalization of the CTT model to address problems related to educational testing were spurred by Kelley (1923, 1927) and Thurstone (1931). By mid-century, the basic tools of classical test theory were widely accessible in Gulliksen (1950) and R. L. Thorndike (1951). However, it was not until the contributions of Melvin Novick and colleagues that CTT was fully axiomatized as a *mathematical or statistical* model (see Novick, 1966; Novick & Lewis, 1967; and chapters 1–11 of Lord & Novick, 1968). (See also Cai et al., this volume, for a discussion of psychometric models.)

CTT is premised on a definition and a thought experiment. Let x_{ij} represent the numeric score on a unique test that has been given to a unique person. The subscript

i indexes the test; the subscript j indexes the person. The observed test score x_{ij} could in principle result from the score assigned to a single test item or task, but it is more commonly assumed to result from the sum of scores assigned to multiple item responses. The fundamental equation of CTT is

$$X_{ij} = \tau_{ij} + E_{ij}. \quad (1)$$

The terms on the right side of Equation 1 are defined as follows. The “true score” τ_{ij} is defined as the numeric value that results from taking the expected value of X_{ij} . An expected value is a mathematical concept and only has meaning with reference to a random variable. This is where the thought experiment comes in. In what sense can an observed test score, x_{ij} , be conceptualized as the realization of a random variable, X_{ij} ? With the help of a little brainwashing.

Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favor of the United Nations; suppose further that after each question we “wash his brains” and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we can then compute the proportion of times Mr. Brown was in favor of the United Nations. (Lazarsfeld, 1959, as cited in Lord & Novick, 1968, p. 29)

The idea here is that if it were possible to observe the results from an infinite number of independent replications of a testing procedure, what we would observe would be a distribution of values. Since the mean of this hypothetical distribution is τ_{ij} and since $E_{ij} \equiv X_{ij} - \tau_{ij}$, it follows that the expected value of E_{ij} will be 0 by construction. Hence, in our thought experiment, the reason that x_{ij} deviates from τ_{ij} is because of some random source of error, which we might think of as good or bad “luck.” Without further elaboration (and generalizability theory, discussed next, was developed to provide such an elaboration), the sources of the good or bad luck are unclear. It could be related to some aspect of the testing procedure, some characteristic of the person being tested, or the interaction between the two that could change across brainwashings. But irrespective of the causes of variability in responses, the quantity of interest is $\sigma^2(E_{ij})$. If this variance is large, it indicates that x_{ij} will be an unreliable measure of τ_{ij} . And because $\sigma^2(E_{ij})$ can differ by person, the same test could conceivably offer a reliable measure of one person’s true score, but an unreliable measure for another.

The statistical challenge of CTT is to devise an approach that can be used to obtain an unbiased estimate of the average of $\sigma^2(E_{ij})$ for a population of test takers. For details on how this challenge is met through the derivation and estimation of a reliability parameter—the proportion of variance in observed scores among a population of test takers that is “real” and not attributable to chance—see Lord and Novick (1968), Haertel (2006), or many other didactic presentations of the CTT linear model available on the Internet. Statistical approaches for estimating a reliability coefficient are all premised on some version of the assumption that it is possible to create a second test that

is “parallel” (according to certain criteria) to the test of interest. This might be done by literally constructing and administering a second test, designed to be parallel to the first. Or, in an approach more commonly taken in the computation of “Cronbach’s alpha,” an existing test is conceptualized as one that could be split into a smaller set of parallel subtests such that the full test can be cast as a composite of these subtests (in the limiting case, a single item could comprise a unique subtest).

There is much that can be done with a suitable estimate for a reliability coefficient in hand (see also Lee & Harris, this volume, for a discussion of reliability). The statistic provides a basis for comparing different tests administered to the same population that not does not depend on the scale of the test; all things equal, if two tests designed to measure the same construct differ in their reliability, the test producing more reliable scores is to be preferred. In a related sense, the general Spearman–Brown formula can be used to estimate the number of test items needed to achieve a desired level of reliability. It can be shown that the best way to maximize reliability with the number of items held constant is to select items of moderate difficulty that are strongly intercorrelated (e.g., Richardson, 1936). Beyond this, the reliability coefficient is properly viewed as a means to the end of estimating a standard error of measurement, useful in quantifying uncertainty about the value of an individual test taker’s true score.

But what is τ_i exactly? There are at least two ways of conceptualizing the meaning of a true score within CTT. In the original formulation of the linear error model by Spearman and Yule near the turn of the 20th century, the observed score on a mental test was split into the actual measure of interest and an increment attributable to errors of observation. This corresponds to what Lord and Novick (1968) described as the “Platonic” conception of a true score as the value of an attribute the test was intended to measure. The true score defined above (i.e., Equation 1) represents an alternative, *operational* conceptualization of the true score as the long-run average that one would observe if it were possible to collect the results from independent replications of the same testing procedure on the same individual (i.e., the brainwashing thought experiment). Now, in a practical sense, this may not seem to matter, since either conceptualization can be used to motivate the same estimate of a reliability coefficient. But each conceptualization may correspond to a different approach to establishing a substantive theory that can predict and explain individual differences for the construct of measurement, and this in turn might influence the approach being taken to design and administer a test instrument.

Generalizability Theory

In discussing the construction of the CTT model, Lord and Novick (1968) had anticipated the extension of the model that was already underway in the form of *generalizability theory* (GT; Cronbach et al., 1972). In a section entitled “What Is Error?” they wrote,

The error random variable and the true score are determined by the experiment, not by some hypothetical state of affairs. . . . The error random variable may have a number of components associated directly with the physical measurement process,

others associated with the environment in which the measurements are taken, and still others associated with temporal changes in the person. Differing degrees of control and randomization of these components define different true scores and hence different residual (error) scores. For each definition of a true score, of course, we have a different error score. (Lord & Novick, 1968, p. 39)

To appreciate what they meant by this, we can translate the “components” to which Lord and Novick alluded into a scenario in which a putative measurement procedure had been envisioned in which a target population of students is being administered a set of test items, and the responses to these test items are to be scored by one or more “raters.” Importantly, the particular set of items that appears on the test, and the raters that score a student’s responses to the items, are not fixed. What Lord and Novick referred to as components would become known as *facets* of a measurement procedure in GT parlance. These facets represent distinct sources of “error” in the sense that they vary at random whenever the measurement procedure is being invoked. For the concept of “at random” to be sensible, under GT anything defined as a facet must come with a corresponding population of “admissible” observations (i.e., all items and raters that would be deemed effectively interchangeable for use on a test). The introduction of facets creates the need for a new thought experiment. Whereas under CTT one must imagine the distribution of scores that would be observed if the same student could be given the same test over and over, we now imagine the distribution of scores we would observe if, following each test administration, we could create a new test, constructed by sampling N_j items and N_k raters at random without replacement from the respective populations of items and raters. As such, the true score for a given student now has a different meaning. It is the average score we would observe if it were possible to repeat a crossed factorial experiment in which the test on each replication is constantly changing as a function of the items and raters being sampled. In GT, this true score is renamed as a student’s *universe score*, since it always has the potential of involving the crossing of samples from more than one population of a measurement facet.

From a modeling perspective, GT can be said to represent an extension of CTT because it serves to essentially decompose error variance into specific sources using analysis of variance techniques. In some sense, it is the computation of these variance components that represents the easier part of GT. The harder part is the conceptualization of the facets of measurement, the populations of admissible observations, and the intended universe of scores over which a generalization is desired. The prioritization of the facets over which a score generalization is desired becomes a critical conceptual move because—in principle—it should be combined with an actual experimental study in which both subjects and facets have been randomly sampled to permit the accurate estimation of the relevant variance components (known in GT as a “G study”). In GT, it is the responsibility of those involved in the design and use of a test to operationalize a universe of facet combinations over which score generalizations can be made.

Many important questions arise in the process. How should one define a population of “admissible” test items? Is the population to be narrowly defined and finite (e.g., all the multiplication problems $m \times n = ?$ where $m, n \in [1, 10]$, as described by Markus & Borsboom, 2013)? Or is it broad and potentially infinite (e.g., all items that could be written by crossing what students are expected to know about mathematical concepts with the practices by which they are expected to invoke them to solve real-world problems)? Should the population of items be considered univariate (e.g., all mathematics content in Grade 5) or multivariate (e.g., the specific Grade 5 content domains of number sense, geometry and prealgebra)? What facets of the measurement procedure are most readily manipulable in an experimental sense? Which facets should be fixed? Which are ignorable? Should the facets be conceptualized as crossed (all items crossed by occasions) or nested (items are unique to a given occasion)? For more detail on this and other aspects of GT, see Cronbach et al., (1971), Shavelson & Webb (1991), and Brennan (2003).

Both CTT and GT provide a basis for homing in on the design of a test that can minimize uncertainty in results due to a particular operationalization of “chance error.” In CTT, because error is that which is orthogonal to a true score in any single administration of a test, it represents something of a black box. Indeed, since the CTT thought experiment only involves replications of the same procedure, it is not clear just why one should expect an observed score to change across hypothetical replications of a test given that all the features of the test itself are fixed. This gets elaborated in GT by more clearly specifying a sampling model for test construction, one in which facets of measurement (e.g., items, raters, occasions) are being drawn from their respective populations. In this sense, GT provides some new conceptual tools that can be used to open the black box. It also provides at least some sense in which CTT can be rescued from the criticism that every test must have a unique true score by introducing the concept of a universe score.

At the same time, the ontological status of a universe score retains some of the same fuzziness as the true score. Is it only a statistical parameter, just more esoterically defined than it was under CTT, or is it intended as a measure of some mental construct within the test taker? Because CTT and GT were being formalized between 1950 and 1970, there can be little question that those contributing to the literature on educational measurement took what can be characterized as either an agnostic position about the existence and interpretation of any underlying construct or, in some cases, a more overtly skeptical position, captured rather nicely by Cureton (1951, p. 641), who wrote,

“Generosity,” “kindness,” “skill,” “proficiency,” and the like are words representing abstract concepts. When we ascribe no more meaning to such a term or construct than the common features of the acts summarized by the word, the term and concept are said to be “operationally defined.” So long as we stick to operational definitions in discussing and thinking about the abstract concepts used in educational measurements, we will be on safe ground. . . . The performer whose average score is high is termed mentally able. We must not say that his high score is due to

his high ability, but if anything, the reverse. We say he has high ability because his performance has yielded a high criterion score. His “ability” is simply a summary statement concerning his actions.

In this way of thinking, a test is conceptualized as a device for taking samples from behaviors reflecting educational objectives that students are expected to have mastered. The construct of measurement is thus defined operationally as some composite of the behaviors that have been formally or informally associated with the content objectives being tested. When the construct being measured by a test is defined operationally by the chain of logic that connects it to empirically observable behaviors, the place to look for substantive theory related to the construct is primarily in the articulation of a replicable process used to articulate a universe of admissible observations (e.g., the blueprints for item design, the protocol for rater training) and draw random samples from this universe.

Item Response Theory

Item response theory (IRT) models were developed to answer questions that the CTT and GT models cannot answer—or at least cannot answer within the context of their mathematical formulation. As we have seen, CTT and GT both revolve around models that were created to quantify the uncertainty in a test instrument that is attributable to random errors in an intended measurement procedure. But to the extent that the linear error models of CTT and GT represent measurement models, they are missing something quite important: scales and units to ensure that the numeric values that result from a measurement procedure are meaningful to those attempting to interpret the result and invariant to differences that arise in the facets of measurement across replications of the procedure. The scale of the test scores that results from a CTT and GT model are locally defined and sample dependent: locally defined in the sense that the scale is a direct function of the scoring rules that have been put in place to transform a student’s set of responses on a test into a numeric value; sample dependent in the sense that any statistic one computes for the test depends on the samples of students and measurement facet conditions. In the absence of a common scale that is invariant to these sampling decisions, it becomes quite challenging to use different tests to monitor trends in performance for a population of students over time, compare differences among groups of students, or make diagnostic inferences about an individual student’s strengths and weaknesses.

IRT offers a possible solution to these challenges because, in principle, an IRT model produces an estimated location for a construct of interest on a scale of measurement that is invariant to the choice of items that comprise a test. And symmetrically, the model can be used to produce estimates of the difficulty of test items that are invariant to the choice of subjects that respond to the items. The cost of these benefits comes in much stronger assumptions than CTT, but these assumptions serve to connect CTT back to the factor analysis tradition that motivated its fundamental equation (Briggs, 2021b). In what follows, we assume the reader is familiar with the basic logistic parameterizations

of IRT models for dichotomously and polytomously scored items. For details, see Baker & Kim (2004), van der Linden (2016), and the many didactic presentations available in introductory textbooks and on the Internet.

A critical assumption made in IRT is that of conditional (local) independence. When conditional independence holds, the existence of a continuous latent variable, θ , typically posited to be unidimensional, is taken as a summary explanation for the intercorrelation among test items. According to the model, if one could restrict a population of test takers to those with identical values of θ , all remaining variability in responses across individuals for any given item would be attributable to chance. When an IRT model shows adequate fit to empirical data, it comes with the very appealing property of parameter invariance. Parameter invariance promises that what one learns about the parameters for a set of test items (e.g., their difficulty and discrimination) does not depend on the ability of the particular sample of people responding to the items, provided that the sample is part of the target population. Through this property, IRT models can be used to equate tests using designs that feature different samples of students taking tests in which some items are common, but others are unique. When this has been accomplished, tests with different true scores can be used to produce comparable estimates on the logit scale of the latent variable θ .

Although IRT retains the same emphasis as CTT and GT on using a quantification of uncertainty about the construct of measurement as a criterion for test design, the notion of measurement error is one that depends on the relative locations of test items and test takers on the underlying logit scale. This can be appreciated by the fact that after fitting an IRT model, one can compute the amount of “information” each item contributes at each possible location of the ability continuum. Technically, information is the reciprocal of the precision with which a statistical parameter has been estimated, a concept introduced by the statistician R. A. Fisher in his development of maximum likelihood estimation. Each item contributes some information about the location of an individual on the ability continuum; the sum of item information functions produces a test information function; and the square root of the inverse of this information produces a conditional standard error of measurement function (see, e.g., Baker, 1985, for a nice introduction). A consequence is that given a population of test items with known (i.e., previously estimated) parameters, unique test forms can be constructed that minimize uncertainty due to measurement error at any targeted location on the underlying scale, provided that sufficiently discriminating items exist at the targeted locations. This is the feature of IRT models that has driven the development of computer-adaptive testing, a technology of dynamic test design that would not be possible (or at least much more difficult to accomplish) using CTT or GT.

The latent variables in IRT models are specified as if they were continuous quantities. In this way, the analysts who invoke such models follow the precedent from the foundational work described earlier by Fechner, Spearman, Thorndike, and Thurstone, among others. Indeed, in linking their derivation of the normal ogive version of the two-parameter IRT model to factor analysis, Lord and Novick (1968)

also assumed that observed binary item responses can be conceptualized as the dichotomization of a unique continuous quantity for each unique test item. It is seldom clear whether such assumptions represent a pragmatic⁴ or realist commitment about latent variables as constructs of measurement. Under a pragmatic commitment, one may define θ to be a continuous quantity only because doing so makes it possible to predict, with sufficient accuracy to be useful, what will happen when a student at any given location of logit scale is presented with an item at any other location on the scale. In contrast, when making a realist commitment, one is staking a claim about the structure of the construct of measurement and why it is that individuals are expected to vary on this construct. This would seem to require additional substantive theorizing about the nature of the construct, as well as its relationship with other constructs.

In their original presentation of IRT as a “latent trait model,” Lord and Novick (1968) clearly embraced a pragmatic approach.

Any theory of latent traits supposes that an individual's behavior can be accounted for, to a substantial degree, by defining certain human characteristics called traits, quantitatively estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict or explain performance in relevant situations. (Lord & Novick, 1968, p. 358)

The socially constructed latent trait is to be interpreted in a metaphorical sense:

It is sufficient that a person behaves as *if* he were in possession of a certain amount of each of a number of relevant traits and that he behave as *if* these amounts substantially determined his behavior.” (Lord & Novick, 1968, p. 358, emphasis added)

In general, Lord and Novick (1968) conceptualized a true score and latent trait as representations of the same behavioral composite, but expressed on a different scale of measurement. This seems to follow from the fact that mathematically, after fitting an IRT model, one needs only to compute a test characteristic curve as the sum of item characteristic curves and then use this function to relate a latent trait to a true score. There would seem to be at least one logical problem with this position, articulated by Borsboom (2005) and Markus and Borsboom (2013). As we have seen, a true score in CTT, and its extension, the universe score in GT, are numerical values that are defined by the choice of items that appear on a given test instrument, irrespective of how those items are selected. But there is nothing that would guarantee that the intercorrelation of these items need be explained by a single underlying latent trait, θ , which is the defining feature of an IRT model. Therefore, if θ is known, one can derive a true score specific to any set of items that fits the IRT model. But the relationship is not symmetric—knowledge of a test's true score is not guaranteed to provide insights into a value of θ that will satisfy the IRT assumption of conditional independence. This would suggest that a true score and a latent trait should be viewed as two fundamentally different concepts.

The Rasch Model

Mathematically, the Rasch model can be viewed as a constrained version of a three-parameter logistic IRT model (3PLM; Birnbaum, 1968) with no lower asymptote and item characteristic curves with common slopes, and in this sense it can simply be referred to as a one-parameter logistic model (1PLM; Thissen, 1982; Thissen & Steinberg, 1986). However, from Rasch's perspective, the requirement of no varying lower asymptotes or discrimination parameters needed to be satisfied to construct an educational measuring instrument that had properties of invariance analogous to a physical measuring instrument. In this sense, the model is often associated with an entirely different data-model paradigm for educational and psychological measurement (Andrich, 2004). Under Lord and Birnbaum's IRT paradigm, the choice of model follows from features of the test data at hand. For example, if preliminary classical item analyses suggest that there is significant variability in the correlations between items and the total score on a test, and if the item format involves selected response options that can be answered correctly by guessing, then it would seem sensible to model test item responses according to the 3PLM. In contrast, under the Rasch paradigm, test items must be designed (or refined) until they can be shown to meet the criteria for a measurement model with certain properties.

The property to which Rasch attributed the greatest importance was something he referred to as specific objectivity, but which we might describe in more general terms as invariance of comparisons, or the (invariant) meaningfulness of a unit of the scale on which measurement results are reported. The basic idea is that when the responses from some population of people and items fit the Rasch model, *differences* in the ability of any two people will not depend on the reference set of items used to facilitate the comparison. Symmetrically, *differences* in the difficulty of any two items will not depend on the reference sample of people. Specific objectivity follows from two important mathematical features of the Rasch model known as separability and sufficiency. For details, see Engelhard (2013), Fischer and Molenaar (1995), and Sijsma and van der Ark (2020).

The Rasch model bears an interesting similarity to another innovation that was taking place outside the context of educational measurement as part of the development of what is known as representational measurement theory.⁵ Of particular relevance is the theory of conjoint measurement (Krantz et al., 1971; Luce & Tukey, 1964). One of its instances, additive conjoint measurement, provides a framework for testing the hypothesis that a variable of interest can be located on a quantitative continuum, and as such, whether the variable can be represented on a scale with interval properties. Conjoint additivity applies when it is possible to transform two variables so that their additive combination forms a third variable without any interactions. Just as Rasch had sought to illustrate his model and its property of specific objectivity by reference to classical physics and Newton's second law of motion (Andrich, 1988; Rasch, 1980, pp. 109–125), Luce and Tukey (1964) had taken a similar tact in illustrating the concept of conjoint measurement.

It has been argued that the Rasch model can be interpreted as a probabilistic version of additive conjoint measurement (Bond & Fox, 2015; Brogden, 1977; Perline et al., 1979). A similar argument cannot be made for other IRT models with more complex item parameterizations. Following this line of reasoning, the sense in which the Rasch model produces a scale with interval properties comes not from any distributional assumptions (i.e., normality) made about the latent variable (i.e., θ), but from the interpretation of individual differences in θ with respect to a designated reference interval of item difficulty. Just as one could observe the same acceleration by maintaining the same ratio of force to mass, under the Rasch model one can observe the same probability of a correct response by maintaining a constant trade-off between person ability and item easiness—as one goes up, the other must go down. The analogy, however, quickly drifts into the same thought experimental territory as CTT, since a person’s “ability,” unlike force or mass, is not readily subject to manipulation. What would seem to be required to further realize the analogy between conjoint measurement theory and the Rasch model would be strong substantive theories about learning that could be used to, at a minimum, lead to the design of items according to factors that are strongly predictive of their difficulty. We will consider an example of this in the section “Measuring Reading Comprehension.” For more on Rasch measurement theory, see Wright (1997), Engelhard (2013), Bond and Fox (2015), Andrich and Marais (2019), Wind and Hua (2022) and Wilson (2023).

Statistics, Psychometrics and Computational Advances

Statistics can be broadly defined as the science of collecting, analyzing, presenting, and interpreting data. It follows that statistics and statistical modeling have a role to play any time a test has been administered to facilitate the collection of data, since these data will invariably be analyzed, presented, and interpreted, and this will be the case whether the test is administered in an educational or a psychological context. The development and use of statistical models in a psychological context defines psychometrics as a subdiscipline of statistics. For example, Jones and Thissen (2006, p. 21) described psychometrics as “the disciplinary home of a set of statistical models and methods that have been developed primarily to summarize, describe, and draw inferences from empirical data collected in psychological research.” If one takes a narrow view by identifying educational measurement solely with the application of the models described in the previous four subsections (which, to be clear, we do not), then the primary difference between psychometrics and educational measurement is the context and purpose for which empirical data have been collected. The statistical models and methods of inductive reasoning found in each are the same.

An important distinguishing feature between the models one finds within CTT and GT relative to the taxonomy of models found in IRT (and here we include the Rasch model within this umbrella) is that IRT models involve assumptions that can be tested and rejected when they appear implausible in the face of empirical data. The origin of CTT predated many of the mathematical formalizations of statistical models

with respect to random variables, though this was later retrofitted to justify some of its central results. In contrast, IRT models were formulated from the outset so that a clear distinction could be made between the parameters of the model and the estimation of these parameters from data. What was already clear by the time of Lord (1980) was that while some of the more complex IRT models seemed suitable as a means of addressing a host of practical problems in educational testing context, they could only be implemented at scale by overcoming some significant computational challenges in the estimation of model parameters.

Briefly, to appreciate the estimation challenge, let the term P_{ij} represent the functional form of an IRT model, where i indexes a test selected-response item ($i = 1, \dots, n$) and j indexes a person taking the test ($j = 1, \dots, N$). Let x_{ij} represent one of two possible scores; $x_{ij} = 1$ represents a correct response, and $x_{ij} = 0$ is an incorrect response. When item and person parameters are known, then P_{ij} returns the probability of a correct response from person j to item i . Following the assumption of conditional independence among items and statistical independence among respondents, a joint likelihood function can be written as

$$Prob(X|\theta) = \prod_{j=1}^N \prod_{i=1}^n P_{ij}^{x_{ij}} (1 - P_{ij})^{1-x_{ij}} \quad (2)$$

The original joint maximum likelihood estimation approach proposed by Birnbaum (1968), and implemented in the earliest computer programs (BICAL for Rasch models; Wright & Mead, 1978; LOGIST for the 3PLM, Wingersky et al., 1982), proceeded by taking the log of Equation 2 and then setting simultaneous estimation equations after taking the partial derivative of the log likelihood function with respect to each item and person parameter. Joint maximum likelihood estimation would then iterate between items and persons, treating one set of parameters as known while solving for the others using the Newton–Raphson technique. The iteration would continue until some convergence criterion was met.

By the early 1980s, a tractable alternative to the joint maximum likelihood estimation approach in the context of models with item discrimination and guessing parameters had been introduced by Bock and Aitkin (1981) that involved integrating over the distribution of person ability and then iterating a solution through use of the expectation–maximization algorithm (Dempster et al., 1977). The details of the development of marginal maximum likelihood estimation are outside the scope of this chapter, but see Bock (1997) for an overview and Baker and Kim (2004) for a comprehensive account. What needs to be appreciated is that the innovation of marginal maximum likelihood estimation also introduced new elements with conceptual significance. Importantly, the marginalization of a likelihood function requires an explicit assumption to be made about the population distribution of θ , which has implications for interpretations about the meaning of probability in an IRT model and for the implied source of measurement error when a person comes in contact with a test item (Holland, 1990). The addition of a population distribution would eventually make it easy to envision new ways that the IRT model can be extended by, for example, expressing

the mean of the person population distribution as a function of other covariates (e.g., “latent regression,” see Mislevy, Johnson, & Muraki, 1992). The marginal maximum likelihood estimation innovation also introduced a Bayesian conceptualization of probability into the estimation of person parameters in the form of a posterior distribution. This formalized the idea that what an IRT model can (and perhaps should) provide is not so much a single measure of an attribute, but a distribution of plausible measures, with the range of these values quantifying the certainty of the measurement procedure.

The advent of the personal computer and the exponential increases in storage and processing power throughout the 1990s and early 2000s made large-scale applications of IRT a reality not only for testing companies, but also even for academic researchers. Highly specialized software proliferated, but also may have obscured the connections between IRT and a broader class of statistical models. By the turn of the 21st century, these connections were being made more transparent because it could be demonstrated that IRT models are special instances of generalized linear and nonlinear mixed models (e.g., McDonald, 1999; De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004), with parameters that are estimable using mainstream commercial software (e.g., SAS, Stata) and even freely available open-source software (e.g., R). Since 2010, new advances that capitalize on a Bayesian conceptualization of probability and Markov chain Monte Carlo estimation have opened the door to the estimation of complex, multidimensional models that were only a glimmer in Frederic Lord’s eyes in 1980 (e.g., Cai, 2010a, 2010b, 2017).

MEASUREMENT AND THE TISM FRAMEWORK

Having traced some of the key developments that have expanded the reach of measurement terminology and practices from the physical sciences to the human sciences, and ultimately into the domain of evaluative decision-making about teaching and learning, we now return to the general definition of measurement we provided at the outset of this chapter.

Measurement is a process that (a) provides information about a property of an object or event in the world, (b) uses instrumentation designed to be sensitive to differences in this property, (c) produces information in the form of values related to a scale or unit for that same property, and (d) is qualified by information about uncertainty.

Four components comprise this definition, and each one corresponds to a core component of the TISM framework that we now introduce and illustrate in detail.

Theory, Instrumentation, Scales and Units, and Modeling

In our view, the process of measurement necessarily involves interactions between four bodies of knowledge and practice, which we refer to as the four central components of measurement:

1. *Substantive theory.* The generation and refinement of theory (e.g., a physical law, a cognitive model) to define the target property or construct (e.g., temperature, reading ability) and to describe the sense in which this property or construct is instantiated in the relevant object (e.g., a vessel of water, a person). A key feature of such theory is that it is amenable to empirical investigation and potential revision.
2. *Instrumentation.* The development of instrumentation (e.g., a thermometer, a reading test) designed to be maximally sensitive to variation in the target property or construct across objects (e.g., different vessels, different people) and/or time and minimally sensitive to everything else.
3. *Scales and units.* The development of scales and potentially units to ensure that the (usually numeric) values that result from measurement will have an interpretation that is meaningful, publicly explainable, and invariant (within specified limits) to the objects being measured, the instruments used to produce the numeric values, and the person who is interpreting the values.
4. *Modeling.* The application of formal models for the purpose of characterizing the uncertainty of the measurement results.

We depict these four components of measurement and their interactions visually in Figure 21.1.

We find the TISM framework helpful as a way of organizing and analyzing the features of measurement processes that serve to secure and justify the claimed meaningfulness

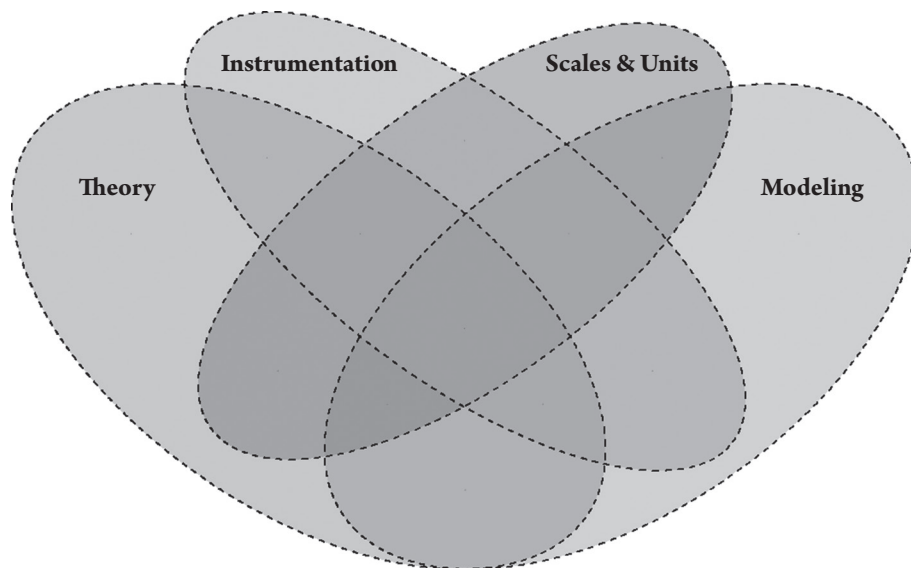


FIGURE 21.1
Central Components of the Process of Measurement

and trustworthiness of measurement results. One thing that the visual representation of our framework as a Venn diagram emphasizes is that the process of measurement involves two- and three-way interactions among these four components and that, ideally, a coherent perspective on measurement places it at the center of a four-way interaction.⁶ From this vantage point, testing in and of itself is simply a form of instrumentation. The extent to which a given instance of testing is an instance of measurement depends on how test design has been informed by substantive theory about the target construct, the approach taken to place test results on an appropriate reporting scale, and the way that models are being used to characterize uncertainty and evaluate hypotheses about the construct.

In what follows, we show how the TISM framework can be used to unpack the features of a measurement process with two concrete examples. To emphasize the point that the framework can be applied to measurement in both the physical sciences and education, we first briefly discuss the application of the TISM framework to the measurement of temperature. We then provide a more detailed treatment in the context of the measurement of reading comprehension as it occurs in two assessment programs in the United States, the National Assessment of Educational Progress and the Smarter Balanced Assessment Consortium.

Measuring Temperature

Temperature measurement, that is, thermometry, is one of the more recent examples of physical measurement, including the historical development of measuring instruments that represented levels of temperature on ordinal (e.g., thermoscopes), interval (e.g., Celsius thermometers), and finally ratio (e.g., Kelvin thermometers) scales, as the theory and practice of thermometry matured. For two accounts that follow these developments in more detail, see Chang (2004) and Sherry (2011).

Substantive Theory

Contemporary thermometry is underpinned by centuries of empirical observations and experimentation, culminating in the development of the mature scientific field of thermodynamics, which is the T of our TISM framework. Thermodynamics consists of definitions, laws, and models related to temperature and other thermodynamic phenomena such as heat and work. One of the foundational empirical observations in establishing the field of thermodynamics, recognized since antiquity, was the systematic relationship between changes in the temperature of a substance and changes in its volume, which in turn laid the foundations for the development of thermoscopes and, later, thermometers. At first, this systematic relationship between temperature and volume was only understood in ordinal terms, and so thermoscopes, typically some kind of vessel or tube in which a substance like water or alcohol could expand or contract, were developed where their scales only reflected ordered differences in temperature. This all changed between the 17th and 19th centuries with the development of the simple and ideal gas laws, which formalized the relationships between temperature,

volume, and pressure, allowing temperature to be understood quantitatively through these interconnections, which in turn supported the development of the first thermometers (e.g., Celsius, Fahrenheit) that were able to measure quantitative differences (i.e., intervals) in temperature.

The understanding of temperature further evolved with the advent of statistical mechanics, a branch of thermodynamics, which defined temperature as the average kinetic energy of the particles in a substance. This theoretical development led to the conception of an absolute zero for temperature, representing the complete absence of such kinetic energy, enabling temperature measurement on a ratio scale. This was first achieved by converting the scales of existing thermometers to the absolute (e.g., Kelvin) scale, but the maturation of thermodynamic theory expanded the comprehensibility of the concept of temperature and diversified its potential measurement approaches. For instance, the Thomson effect (Thomson, 1851), which describes the relationship between temperature and electrical voltage, provided the theoretical foundation for many digital thermometers that we use in the early 21st century.

Instrumentation

The I in the TISM framework applied to thermometry represents the theoretical and practical knowledge of thermal properties like expansion and conductivity in various materials. These properties are crucial for constructing instruments that can accurately sense temperature changes while minimizing non-temperature-related influences. In a liquid-in-glass thermometer, for instance, it is important to know the liquid's rate of thermal expansion across the thermometer's intended range of use. If this rate is established to be linear across the range, then the expansion of the liquid in the thermometer can be interpreted as reflecting quantitative differences in temperature rather than just ordinal differences. This kind of material knowledge further enabled the technological progression from the early water and alcohol thermoscopes to mercury thermometers that possess such linear thermal expansion.

Creating a liquid-in-glass thermometer also requires specific thermal properties in the glass itself—it must conduct heat to the liquid but show little to no thermal expansion, ensuring that only the liquid in the tube changes volume. Precise engineering is also needed to maintain uniform dimensions across the tube's length. Once these kinds of instrumental requirements are met, the thermometer can systematically transduce quantitative variation in temperature into variation in spatial extent by way of the liquid's volumetric expansion or contraction. These spatial differences can then be assigned temperature values according to the instrument's scale.

Scales and Units

The S of the TISM framework applied to thermometry refers to the theory and practices involved in the establishment of the measurement scales (e.g., Fahrenheit, Celsius,

Kelvin) and ensuring their comparability and convertibility. It also involves the careful calibration of the thermometer to establish its accuracy and precision. For a liquid-in-glass thermometer, calibration entails defining two fixed points on the scale (e.g., the freezing and boiling points of water) and setting equal-interval gradations between them. This, in turn, determines the precision of the scale via the number of gradations between the fixed points.

Converting values between the different scales is made feasible by understanding the differences in the assigned values to the fixed points, as well as their precision. For example, the freezing point of water is approximately 0 degrees Celsius and 32 degrees Fahrenheit, while the boiling temperature of water is approximately 100 degrees Celsius and 212 degrees Fahrenheit, so converting from Fahrenheit to Celsius involves subtracting 32 from the value and multiplying it by $100/180$ (simplified as $5/9$). Using this formula, one can move back and forth between the degrees Celsius and Fahrenheit units while maintaining the invariance of the quantitative temperature differences the values represent.

This primary concern with invariance extends from conversion between scale values to the definition and refinement of the units of measurement over time. Taking the degree Celsius unit as an example, improvements in thermometric theory and technology led to the discovery that the fixed points used in the definition were not invariant across different levels of atmospheric pressure, and so the unit definition was revised to overcome this shortcoming by including points that were defined in terms of both pressure and temperature. Most important for the progression of thermometry, the degree Kelvin was officially adopted as the international unit of temperature in the mid-20th century, which included the absolute zero temperature value as a fixed point in its definition, thereby entrenching thermometry as ratio-level measurement. This unit, now just known as the Kelvin, has continued to be refined to the present time.

Modeling

Finally, the M of our TISM framework includes the various kinds of formal models that are essential to thermometry through, for example, scientific models such as the ideal-gas laws, instrumental models such as the schematic for a liquid-in-glass thermometer, calibration models such as the thermal apparatus used to calibrate the fixed points, and the statistical models that allow us to evaluate the uncertainty of the measurement values. It is the latter that we will focus on given the importance of measurement uncertainty in any specific measurement application.

Measurement uncertainty is generally conceived of in terms of error that has a systematic (i.e., predictable and constant) effect on the values across repeated measurements, versus error that is random (i.e., unknown and unpredictable) in its effect across these values (Bich et al., 2006). In thermometry, systematic errors are typically quantified with respect to inconsistencies between the scientific, instrumental, and/or calibration models and the actual measurement instrument and context, whereas random error

is quantified in terms of a statistical model (e.g., the normal distribution) applied to repeated measurement values.

Taking liquid-in-glass thermometry as an example, systematic error may occur when a thermometer's bulb permanently expands relative to its calibration volume, introducing a systematic negative error in the measurement values. Simultaneously, random error may arise from small and unsystematic differences in the time given to the thermometer to adjust to a thermal phenomenon (e.g., the temperature of a specific liquid) across repeated measurements. While an estimate of the magnitude of this systematic error may be utilized to establish a correction factor for the thermometer's measurement results, it may also be incorporated into the overall uncertainty estimate along with the random error magnitude, typically represented by the standard deviation of repeated measurements. This overall uncertainty estimate may then be used to qualify the thermometer's measurement results.

From Thermometry to Educational Measurement

The use of thermometry to illustrate the four central components of a measurement process might seem to set a rather high bar for the measurement of nonphysical properties (i.e., constructs) to cross. Notice how each component depends on the other. Instrumentation is informed by theory even as theory can only be tested with provisional instrumentation. Substantive theory is formulated with respect to definition and laws, and these are expressed as mathematical models, which are only approximations. Models are also used to quantify uncertainty, but this quantification depends on choices made in regard to scales and units. Finding an example in which the four components can be as seamlessly integrated in educational measurement is indeed a tall order, in particular, because our theoretical understanding of educational and psychological constructs is nowhere near as strong. Psychological laws are few and far between (if they exist at all), and unlike physical properties like temperature, where there was a convergence over time on its current thermodynamic definition that seems unlikely to change, the way that constructs like intelligence, reading ability, and self-efficacy are perceived and conceived can and do change, perhaps even dramatically, over time.

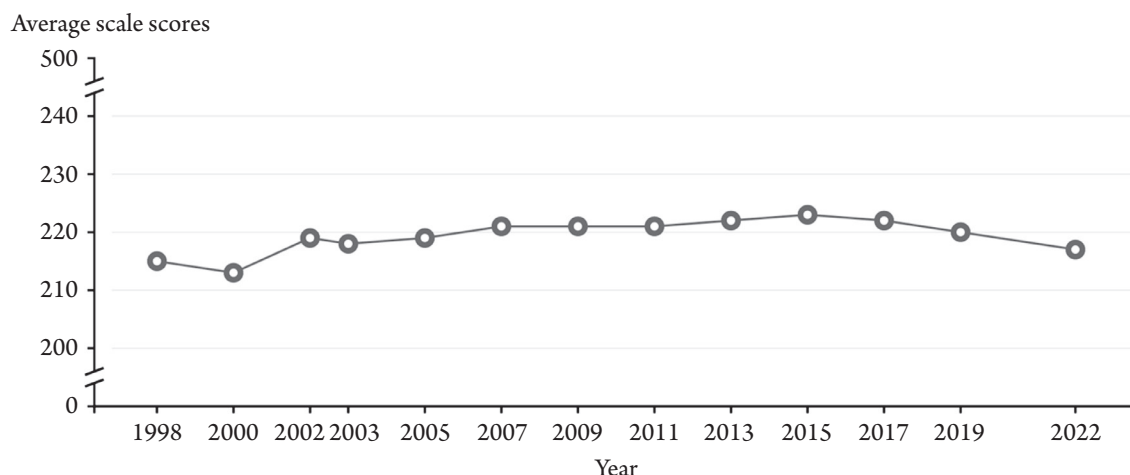
At the same time, many of the pioneers in applying the terminology and methods of physical measurement to the domains of psychology and education quite consciously used the success of thermometry as a reason for optimism (Briggs, 2021b; Galton, 1877). It took centuries before thermometry advanced to the point where the observed effects of differences in temperature could be connected to thermodynamic principles and ultimately measured on a ratio scale. The attempt to do something analogous in the measurement of reading ability is only a little over 100 years old, so who knows what advances might be possible in another 100 years. And as we will show in the next section, contemporary methods used to measure reading comprehension do involve all four of the central components of our framework—even if the coherence of their integration may not always be given as much attention as is warranted.

Measuring Reading Comprehension

We now apply the same TISM framework to an example from educational measurement, namely, that of reading ability as it takes place in two contemporary large-scale assessment programs in the United States. Developing the ability to comprehend text has been a long-standing objective of nearly all systems of public education. In many countries it is one of the two subject domains (along with mathematics) within which the ability of children is ostensibly measured on a periodic basis between the ages of about 6 and 18. As we will show, the measurement of reading ability depends on the successful integration of bodies of knowledge and practice related to theory, instrumentation, scales and units, and modeling. To make this illustration more concrete, we will draw from the National Assessment of Educational Progress (NAEP) and the Smarter Balanced Assessment Consortium. We begin with a brief overview of each assessment program and the salient distinctions in their purposes and contexts for producing measures of reading comprehension.

In the United States, NAEP, known as “the nation’s report card,” (<https://www.nationsreportcard.gov/>) is considered the premier large-scale assessment program designed for the purpose of monitoring trends in student achievement. NAEP tests in reading and mathematics are given every 2 years to nationally representative cohorts of American students at three grade levels (Grades 4, 8, and 12). These tests have been administered in the United States on a periodic basis since 1970, and their results are used to track the progress of successive cohorts of American students. Although individual students are tested, no individual scores are reported. Instead, because the focus is on making inferences about aggregate trends for specific populations and subpopulations of students (e.g., the nation, states, large cities, demographic groups), a great deal of attention is devoted to the sampling of students, and the time requirements for the test are minimized by administering to each student a relatively small sample from the population of available test items. The national trend in NAEP Reading performance of Grade 4 students from 1998 to 2022 is depicted in Figure 21.2, where the horizontal axis represents a year in which the test was administered and the vertical axis represents one of two scales on which test performance is regularly reported.

Smarter Balanced consists of multiple entities (13 states, along with the U.S. Virgin Islands and the Bureau of Indian Education at the time this chapter was written) that administer common assessments to students in American public schools in Grades 3–8 and 11, near the end of each academic school year. We focus here on the Smarter Balanced assessment of English language arts (ELA), within which reading comprehension is identified as one of four “claims” about student abilities that scores on subsets of items on the test are intended to support. An overtly stated purpose of the test is to “accurately describe both student achievement (how much students know at the end of the year) and student growth (how much students have improved since the previous year) to inform program evaluation and school, district, and state accountability systems.”⁷ Even more generally, the test results are meant to be interpreted as an

**FIGURE 21.2**

Trends in National Assessment of Educational Progress Reading Comprehension for Grade 4 Students From 1998 to 2022

Note. Data are from the U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress, 1998, 2000, 2002, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017, 2019, 2022 Reading Assessments.

indication of student progress toward “college and career readiness” by the time they complete high school. Because the assessment is administered using a computerized adaptive test design, Smarter Balanced argues that its tests provide for “efficient and precise” measurement. Unlike NAEP, measurement results are reported for individual students, though they are also aggregated to examine trends across classrooms, schools, and school districts.

An example of a student-specific Smarter Balanced score report for ELA/Literacy tailored to the state of California is shown in Figure 21.3. Notice that reading comprehension measurement results are reported on an ordinal scale consisting of three discrete categories (*Below Standard*, *Near Standard*, and *Above Standard*). In contrast, the overall ELA/Literacy assessment results are reported, like NAEP, on two different scales, one of which is at least implied to have interval properties (with scores ranging from 2131 to 2663) and another resulting from a transformation of this scale into ordered levels.

Substantive Theory

How do people learn to read? Although reading itself has been around in some form since the dawn of human civilization, formal research and scholarship aimed at producing theories that explain how we learn to read is relatively new and still evolving. Although the lines that divide them are sometimes blurry, *learning theories* can be distinguished by the attention they place on behavioral, cognitive, social, and cultural factors involved in the process of developing knowledge and skill.

In general, *behavioral* learning theories focus attention on decomposing the act of reading comprehension into its constituent skills and then identifying the environmental conditions that can be plausibly manipulated to facilitate mastery of these skills

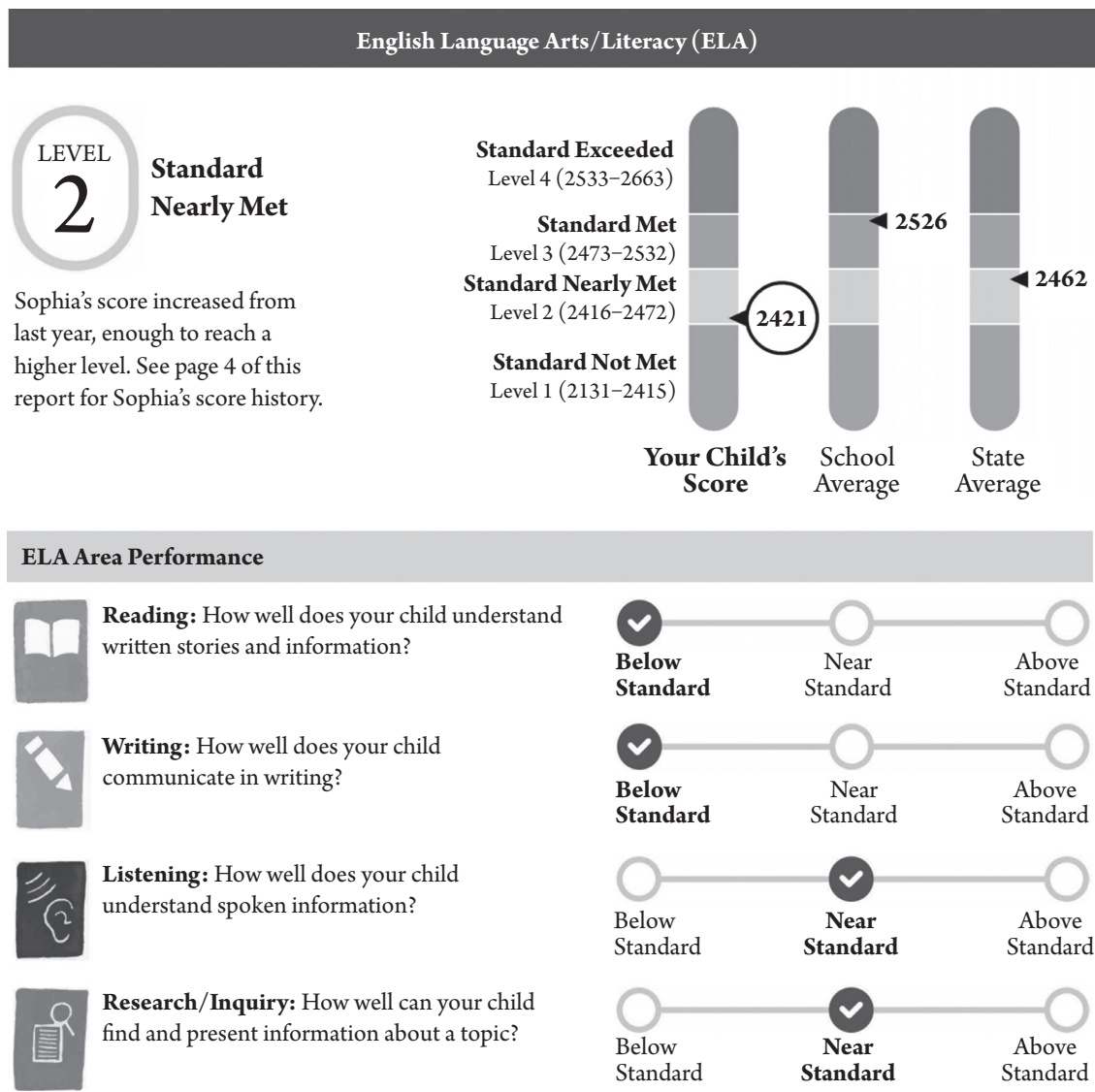


FIGURE 21.3
Mock Report of Student Results From Smarter Balanced English Language Arts/Literacy Assessment
Note. Data are from the California Department of Education website, retrieved June 25, 2020. <http://www.caaspp.org/rsc/pdfs/Smarter-Balanced-Grade-4-Sample-Student-Score-Report-English.2019-20.pdf>

(Alexander & Fox, 2004). A defining feature of the behaviorism predominant in the United States through the mid-20th century was a strong commitment to empirical observation and an avoidance of reference to nondirectly observable entities (such as the mind) and processes (such as cognition) wherever possible (e.g., Skinner, 1938; E. L. Thorndike, 1905; Watson, 1913). In contrast, *cognitive* theories of learning focus on the process of reading comprehension as it is believed to take place within the mind of an individual reader. Notable features of cognitive theories of learning are the focus on the development of knowledge and skill as a learner moves from “novice” to “expert” over

time and on the active role individuals play in constructing knowledge over the course of this development. Overlapping with cognitive theories, *sociocognitive* (and *sociocultural*) theories are premised on the idea that learning, especially in school settings, is an intrinsically social activity that takes place within a cultural context. Sociocognitive and sociocultural theories of learning focus on the roles of knowledgeable others (i.e., teachers and peers) in supporting a developing learner through the use of scaffolding (e.g., Vygotsky, 1978) and the value of drawing on resources from a student's community and cultural context to foster engagement and agency with learning (e.g., Lave & Wenger, 1991). For two important consensus overviews of modern theories of learning and their research base, see National Research Council (2000) and National Academies of Sciences, Engineering, and Medicine (2018).

In a report chapter commissioned by the National Academy of Education, Kendeou (2020, p. 72) described reading comprehension as

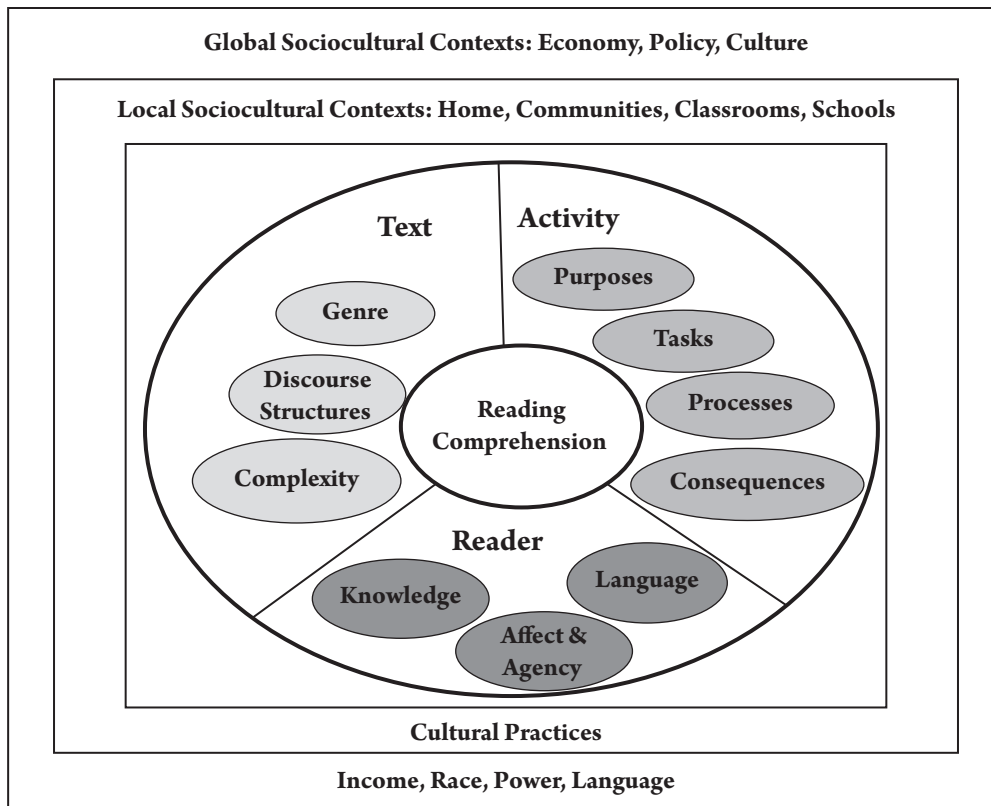
among the most complex of human activities. It involves processing words, connecting words using rules of syntax to understand sentences (Perfetti & Stafura, 2014), integrating meaning across sentences, drawing on relevant knowledge, generating inferences, identifying the structure of the text, and taking into consideration the authors' goals and motives (Graesser, 2015). The end product is a mental representation, what has been termed the "situation model" (Kintsch & van Dijk, 1978), which reflects the overall meaning of the text.

This conceptualization is consistent with an earlier and influential publication by the RAND Reading Study Group in 2002. The panel of reading scholars behind this report defined reading comprehension as the "process of simultaneously extracting and constructing meaning through interaction and involvement with written language (p. 11)." They further proposed that comprehension can be understood with respect to three elements: the reader who is comprehending, the text that is to be comprehended, and the activity for which comprehension is required, all of which are shaped by a larger sociocultural context. For a visual illustration of the key elements of the theory as taken up by a group of national experts asked to develop a new framework that would guide the design of the 2026 NAEP Reading tests, see Figure 21.4.

How does theory inform the measurement of reading comprehension? The 2026 NAEP Reading Framework, adopted by the National Assessment Governing Board (NAGB) in August 2021, incorporates the four central elements of a general sociocultural model of reading comprehension (readers, texts, activities, and sociocultural context) as follows (NAGB, 2021, p. 10):

Reading comprehension is making meaning with text, a complex process shaped by many factors, including readers' abilities to:

- engage with text in print and multimodal forms;
- employ personal resources including foundational reading skills, language, knowledge, and motivations; and
- extract, construct, integrate, critique, and apply meaning in activities across a range of social and cultural contexts.

**FIGURE 21.4**

Visual Representation of a Sociocultural Theory of Reading Comprehension From the Public Comment Draft of the Reading Framework for the 2026 National Assessment of Educational Progress

As illustrated in Figure 21.4, reading is to be understood as a purpose-driven activity in which comprehension depends on the genre, complexity, and discourse structure of the text, as well as the knowledge, affect, agency, and language skills of the reader. A key component of this sociocultural theory is the idea of reading comprehension as a sociohistorically contingent activity, which is greatly influenced by the “scaffolds” available to students at a given place and time (which for most U.S. students would now include the Internet, for example, and may increasingly include tools based on artificial intelligence). This highlights the role of students’ metacognitive attributes in reading comprehension to know when and how to make use of such scaffolds.

Instrumentation

Historically, instruments designed to help make inferences about students’ reading comprehension have come in the form of collections of passages of text to read and related questions administered either orally, on paper, or in a digital environment. The need for guidelines with respect to the design of such measuring instruments was

one of the motivations for both the publication of *Educational Measurement* and the formation of NCME. One of the major recent developments in this body of knowledge and practice is the shift from paper-and-pencil to digital administration (e.g., via a computer monitor or tablet). From its inception in 2015, Smarter Balanced has administered its tests via a digital interface; NAEP transitioned to a digital administration of reading tests in 2017. This has led to a proliferation of possible item formats that can be efficiently administered through so-called technology-enhanced items (e.g., a student can be asked to highlight the text in a passage that provides the best answer to a question). It also suggests new possibilities in making the interactions between students and test items more dynamic through the use of scenario-based tasks⁸ and potentially more attentive to the different elements involved in a sociocultural theory of reading (as summarized in the previous section).

The design of instrumentation inevitably requires reference to substantive theory. For example, a liquid-in-glass thermometer instantiates a *transduction principle*, whereby variability in temperature causes variability in the expansion of a thermometric fluid, based on the laws of thermodynamics. Can we similarly argue that variability in the ability of, say, fourth-grade students to comprehend text can be transduced into variability in their item responses? A key question to be posed along these lines would be, On what basis would we hypothesize that items with a given set of features are effective transducers⁹ of reading comprehension? One attempt to answer this question seems to be implicit when professional item writers work with “task models” (Mislevy & Haertel, 2006) that connect prospective test items to unique content standards (e.g., describing the central ideas and key details of a written passage, defining word meanings). These task models can also provide a basis for hypotheses about items that should be easier or harder to answer correctly as a function of the complexity of the cognitive process involved. For example, after reading a passage, items that require students to locate and recall information might be predicted to be easier, on average, than items that require them to integrate and interpret information or to analyze and evaluate it. This becomes a theory with testable implications for explaining the variability in item difficulty estimates as a function of task model features (see, e.g., Gorin, 2006).

On the Smarter Balanced and NAEP Reading assessments, a connection between theory and instrumentation is most visible through the specification of *achievement-level descriptors*, which are designed to assist item writers in the task of writing items that would distinguish between “achievement levels.” To be more concrete, the task design factors that are posited to distinguish achievement levels on NAEP and Smarter Balanced come from two sources: differences in the complexity of the text that a test taker is given to read and differences in the complexity of the questions that are posed about the text after it has been read. Relative to the sociocultural theory of reading comprehension, what is reflected in the Smarter Balanced and NAEP achievement-level descriptors are implicit cognitive models pertaining to the interaction of a reader and text.

Scales and Units

The bodies of knowledge and practices here are specific to methods for ensuring that measurement results are interpretable by everyone relative to the same standard and comparable (i.e., invariant) irrespective of the occasion of measurement, the specific instrument used, and the person tested. In educational measurement, this encompasses methods that have been invented and developed to scale, link, and equate tests (for a variety of perspectives, cf. Angoff, 1971; Bond & Fox, 2015; Briggs, 2021a; Holland & Dorans, 2006; Kolen & Brennan, 2004). Many methods of scaling applied to educational tests are premised on a desire to “encourage accurate interpretations of test scores and discourage improper score interpretations” by “incorporating normative, score precision, or content information into the score scales” (Kolen, 2006, p. 156). However, there are other methods that are premised on a desire to place both students and items on a common scale with ratio or interval properties and a unit analogous to those found in the physical sciences (Briggs, 2019, 2021a, 2021b; Sijtsma, 2012; Rasch, 1980; Thurstone, 1925).

The reporting scales for reading comprehension on Smarter Balanced and NAEP are quite different (recall Figures 21.2 and 21.3). As previously noted, the Smarter Balanced scale for reading comprehension consists of three discrete categories. In contrast, the NAEP test results are presented on two different scales, one with a theoretical range from 0 to 500 in integer units and the other in terms of four discrete achievement levels. There is a direct relationship between the two NAEP scales because the thresholds for the discrete categories of achievement levels are defined by engaging stakeholders in one of many possible “standard-setting” procedures (see Hambleton & Pitoniak, 2006, Cizek & Bunch, 2007, and Ferrara et al., this volume for overviews; for an interesting design-based approach, see Lewis & Cook, 2020). A similar correspondence exists for the two scales provided for the Smarter Balanced ELA/Literacy test, for which reading comprehension represents a subdomain. On this test, student scores are reported on a common scale that ranges from a low of 2292 in Grade 3 to a high of 2678 in Grade 11 and also with respect to four achievement levels.

What are the properties of the scales used to report measurement results from the NAEP and Smarter Balanced tests? The reporting of results in terms of ordered discrete categories implies a scale with ordinal properties. The interpretation of the finer grained scales that Smarter Balanced and NAEP use to equate tests over time and grades seems more equivocal. Zwick (1992) argued that in the NAEP context, the finer grained scale should also only be given an ordinal interpretation, and as such, statistical comparisons of groups on the scale should be restricted to those that are invariant to all admissible monotonic scale transformations. However, both NAEP and Smarter Balanced scores are frequently used for statistical comparisons of means and variances that imply that the underlying scale has at least interval properties. And the technical documentation for each testing program does not address scale properties. We will return to this issue shortly in the context of the challenge inherent to evaluating magnitude, but for now we note that just as it is difficult to consider choices in regard to instrumentation without

regard to substantive theory about the construct, it can be problematic to separate the interpretation of scales and their units from commitments about theory, as well as from choices about instrumentation and the modeling of item responses.

Modeling

With a few notable exceptions, most large-scale educational assessments invoke the models of IRT (Lord & Novick, 1968; van der Linden, 2016; van der Linden & Hambleton, 1997). As described in earlier in our section on IRT, this can have several practical benefits, provided the model adequately fits the data. First, IRT models person and item parameters in terms of a common numeric continuum. Second, the estimates of these parameters will be invariant up to a linear transformation. Thus, within limits, the estimates will not be biased by the sample of persons or items used for their calibration. These features of IRT make it convenient as a tool for linking and equating different test forms within and across test administrations (as in NAEP) or within the context of a computer-adaptive test administration (as in Smarter Balanced). A third important benefit is the utility IRT provides to quantify uncertainty about measures through test information functions and their inverse, standard error of measurement functions. These functions can be used to help design test instruments that minimize at least some forms of measurement error at targeted locations on the modeled continuum.

This flexibility and generality of IRT modeling is especially important for NAEP. Student responses to NAEP test items are modeled using a combination of two fairly conventional IRT models first introduced in the 1960s, the 3PLM and the graded response model (Samejima, 2010). However, as noted earlier, the unit of analysis on NAEP tests is not the individual student, but a population and subpopulations defined by geographic, demographic, and socioeconomic considerations. As a result, the complete statistical model applied to NAEP tests nests a model for item responses within a model for the population from which a given individual student has been sampled. Although the mathematical details of how NAEP generates aggregate estimates of reading ability from this model through a combination of latent regression and the plausible values drawn from each test taker's posterior distribution is outside the scope of this chapter (for details, see Mislevy, 1984; Mislevy, Johnson, & Muraki, 1992; Mislevy, Beaton, et al., 1992; Oranje & Kolstad, 2019), these details arguably show off an educational measurement success story when it comes to the use of modeling (and probability theory) to solve numerous challenges related to instrumentation in a large-scale assessment program.

The modeling of Smarter Balanced ELA/Literacy tests makes for an interesting contrast. The selected response items on the Smarter Balanced test are modeled using the 2PLM (Birnbaum, 1968) instead of the 3PLM. Does this mean that selected response items on the Smarter Balanced test do not induce guessing? Or has the decision been made for more pragmatic reasons (e.g., guessing is thought to occur, but there is no sufficiently defensible method of modeling its effect on the probability of successfully responding to items)? Another key difference is that the Smarter Balanced tests involve a census rather than a sample of students taking the tests within participating states, and each student

spends much more time taking the test. So, there is not the same obvious need for a population model to accommodate sampling issues. The modeling of the Smarter Balanced tests arguably provides another example of an educational measurement success story when it comes to the interaction between instrumentation and modeling. Specifically, this is the ability to administer test items that are “adaptive” in the sense that the difficulty of a set of reading comprehension items a student receives depends on their performance on a previous set (e.g., Chang, 2015; Meijer & Nering, 1999).

CHALLENGES AND DEBATES IN EDUCATIONAL MEASUREMENT

In the previous section, we introduced the TISM framework and illustrated its four central components in the context of measuring temperature (physical science) and measuring reading comprehension (education). In both instances, the components interact. Advances in our theoretical understanding of a property or construct can change the approaches we take in the design of instrumentation and vice versa. Choices regarding instrumentation also depend on the results from the modeling efforts taken with an eye toward maximizing the consistency of results across replications of the procedure. The construction of a scale for reporting measurement results can depend on the identification of one or more invariant reference standards, and the choice of unit may interact with the approach taken to quantify the amount of chance error to be expected anytime a measurement is taken.

There are unique challenges involved in the measurement of educational (and psychological) constructs, challenges that collectively can make the interpretation of educational measures far more equivocal than the interpretation of most physical measures. In what follows, we describe these challenges, and we turn to the TISM framework to help clarify their nature.

What’s Validity Got to Do With It?

Discussions about the quality and trustworthiness of educational tests and psychological measures often invoke the concept of validity, described by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014, p. 11) as “the most fundamental consideration in developing and evaluating tests” in the human sciences. Although a complete treatment of either the history or the philosophical underpinnings of validity is well beyond the scope of this chapter (for recent comprehensive treatments, see, e.g., Markus & Borsboom, 2013; Newton & Shaw, 2012; Slaney, 2017) and Lane and Marion (this volume) aim to give a contemporary perspective, there is one question we think is worth addressing here: How, exactly, does the concept of validity relate to that of measurement? When we talk about the authority of measurement results, as we do in this chapter, are we really just discussing validity?

Part of the difficulty in addressing this issue is that, as with measurement, there is no universally accepted definition of validity, despite wide agreement regarding its

importance. A starting point might be to say that a measuring instrument's validity is simply the extent to which it measures what it claims to measure: This is arguably the oldest formal definition of validity and may still be the most common understanding of the term across the social sciences in general (see, e.g., Newton & Shaw, 2012). However, at least in the context of educational testing in North America, most commonly encountered definitions of validity are broadly consistent with a perspective often credited to Samuel Messick (1989) and codified by the aforementioned *Standards* (AERA et al., 2014), that is, that validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). This perspective is largely consistent with an "argument-based" approach to validation developed by scholars such as Cronbach (1971), Kane (1992, 2006, 2013), and Mislevy (2018). This approach shifts the focus of validation efforts away from demonstrations of measurement per se and instead toward pragmatic, context-specific arguments for the utility of tests, tailored for specific audiences and circumstances. According to this perspective, the activity of validation is seen as open-ended, with any given test regarded as, at most, provisionally valid insofar as the available argumentation and evidence are found to be convincing by a relevant community of observers and in reference to a specific, socio-historically embedded set of interpretations and uses.

As has been observed elsewhere (e.g., Borsboom et al., 2009; Maul, 2013; Michell, 2009; Slaney, 2017), the gradual shift in focus away from the question of whether a test measures what it claims to measure and toward context-specific argumentation for the adequacy and appropriateness of tests for particular goals has had the result that it is often not obvious how, or even if, contemporary perspectives on validity and validation relate to the concept of measurement. Within the mainstream literature on validity, one would be hard-pressed to find serious engagement with the topic of measurement, even in authoritative sources such as the *Standards* (in which the term *measurement* is frequently used but never defined). At least in part, this disconnect may simply be a function of the fact that the literature on validation—and in particular, each of the editions of the *Standards*, dating back to 1966—has largely focused on the roles that tests play in education (including both large-scale standardized achievement tests and smaller scale tests such as those designed primarily for pedagogical purposes) and, to some extent, in other domains in which tests are regularly used to make decisions about individuals and groups, such as in industrial/organizational settings and counseling and clinical psychology. This is contrasted with settings in which tests are intended more explicitly as scientific measures, such as psychophysics and other areas of basic research. Further, while the contemporary literature on validation provides broad guidance on the aims and structure of the relevant arguments and the kinds of evidence that might be considered relevant (e.g., the five categories of evidence for validity discussed in the *Standards*, AERA et al., 2014), for the most part it does not aim to provide concrete guidance on what specific forms of evidence would be necessary to defend specific kinds of interpretations and uses, instead leaving such decisions to the relevant communities of stakeholders. That is, the emphasis on contextualized argumentation in support of specific, here-and-now decision-making and on the criterion of success

being not truth but the collective judgment of a relevant community arguably renders this conception of validity essentially legalistic rather than scientific (Borsboom, 2012). In contrast, other recent scholarship on validity, largely coming from outside education (and also outside North America) has argued in favor of understanding the semantics of validity in terms of measurement. Borsboom and colleagues have developed an account of validity that could be regarded as an extension of the earliest understanding of the term (i.e., a test is valid if it measures what it claims to measure): specifically, “a test is a valid measure of an attribute if (a) the attribute exists and (b) variation in the attribute causes variation in the outcomes of the test” (Borsboom et al., 2004, p. 1061). This perspective emphasizes that whether or not a test is valid as a measure of an attribute (or property) is a claim about the actual state of affairs in the world and is thus independent of the evidence available at any given time, or the extent to which that evidence is found to be persuasive by any given community of observers (see also Borsboom et al., 2009; Cizek, 2020). Stated alternatively, in apparent contrast to the Messickian perspective, it is possible to be mistaken about the validity of an instrument—a group of stakeholders may make an integrated and evaluative judgment that a given test is adequate and appropriate for the purpose of measuring a particular attribute, but may simply be wrong, just as it is possible to be wrong about any other belief formed about the state of affairs in the world.

From this perspective, the critical issue of validity—at least as it pertains to measurement—is the causal relationship between (inter- and/or intrapersonal) variation in the target construct and variation in the outcomes of the testing procedure that serves as the measuring instrument. In the vocabulary of the TISM framework for measurement described in this chapter, this could be seen as a focus on the transduction between (variation in) the construct and (variation in) the results of the application of an instrument, and thus on the theory and instrumentation components of the TISM framework. A possible point of contention seems to arise from the fact that this is a theory of validity and not of validation—that is, Borsboom and colleagues’ body of work on validity does not aim to offer explicit advice on how to go about establishing validity, but instead focuses just on clarifying the semantics of validity claims—but a fairly obvious corollary of this perspective is an emphasis on the importance of articulating and ultimately testing strong cognitive theories explaining the way in which instruments instantiate relationships between the constructs they intend to measure and the test scores or other results of their application. In this way, the emphasis of Borsboom and colleagues’ perspective is broadly consistent with what we perceive to be a trend in at least some major recent works on assessment design, such as the influential report *Knowing What Students Know* (National Research Council, 2001), which foregrounded the alignment of cognitive theory and assessment design and, in turn, influenced frameworks for the construction of instruments such as the Berkeley Evaluation and Assessment Research Assessment System (e.g., Wilson, 2023) and evidence-centered design (ECD; e.g., Mislevy, & Haertel, 2006). Additionally, it could also be argued that similar or overlapping concerns have motivated a good deal of recent work in psychometric theory, which has emphasized the need for a connection between statistical models with cognitive theory and test design

(e.g., De Boeck & Wilson, 2004; Embretson, 2010; Embretson & Gorin, 2001; Leighton & Gierl, 2007; Mislevy, 2018; Sijtsma & van der Ark, 2020). Using the language of the TISM framework, these trends may indicate a greater interest in the interaction between the modeling component with both the theory and the instrumentation components than has been found in previous decades.

Another measurement-focused perspective on validity has been offered by Joel Michell (e.g., 2009), who argued that a great deal of the literature on validity and validation seems to be premised on a misunderstanding of measurement and, as a result, fails to test the most critical relevant hypotheses, with the conclusion being that many or perhaps even all claims to measurement in education and the human sciences are at best premature. This sobering conclusion is based on a decades-long exploration of the history and philosophy of measurement (to name just a few key sources, see Michell, 1997, 1999, 2005, 2008); while any attempt to succinctly summarize this body of work will obviously miss much of its nuance and depth, the basic premise of Michell's argument (as briefly introduced in the section "Measurement in the Physical Sciences") is that measurement, like scientific inquiry in general, is an activity aimed at discovery of facts about the true state of affairs in the world (as opposed to an act of pure construction, numerical assignment, etc.). More specifically, Michell argued that measurement is historically understood in scientific contexts as being aimed at the discovery of ratios of quantities of a property, usually relative to a unit of that property. (A statement like "Derek is 50 years old" can be understood as "the ratio of Derek's age to the unit of a year is 50.") Michell's perspective thus aligns with that of Borsboom and colleagues in its emphasis that whether a measurement claim is valid is a function of the state of affairs in the world rather than the judgment of any particular group of stakeholders, and it adds an additional specification regarding the output taking the form of estimated ratios of quantities. One corollary of this premise is that processes that lead to results reported on what Stevens (1946) referred to as nominal or ordinal scales of measurement could not be considered actual instances of measurement (just as thermoscopes are not measuring instruments, but thermometers at least potentially are). Another corollary of this premise is that a key task facing anyone interested in measuring an educational construct is to figure out whether the construct is, in fact, quantitatively structured—that is, whether differences in magnitudes of the construct meaningfully admit of ratio-level comparisons—which is an empirical question, to which the answer might in fact be "No," in which case it is not possible even in principle to measure it (though other modes of assessment of the construct may still be possible, such as purely qualitative assessment, or classification, or seriation). If an instrument is constructed to measure a given construct without the due diligence of checking whether it is quantitative (which further presupposes that the property has been defined with a sufficient degree of clarity and precision to allow such a question to be meaningfully asked of it), and if results from the application of this instrument are presented in quantitative terms, presupposing the meaningfulness of units—for example, "This student's reading comprehension ability is 550 units, which is 150

more than last year and 1.5 times the average year-to-year-change”—such results may simply not be credible, and presenting them as results of genuine measurement processes could be considered duplicitous.

It seems fair to say that, for the most part, Michell's line of critique has not led to a significant re-evaluation of mainstream practices in educational measurement. We might speculate as to some of the contributing factors: First, the critique is nuanced and may be challenging to understand without substantial background in the history and philosophy of science, which is likely not a point of emphasis in the training of many testing professionals and social scientists; second, for the most part, the critique has not been developed in a way that makes it clear exactly what consequences it has for mainstream practice, particularly in educational testing (though see, e.g., Briggs, 2013; Briggs & Domingue, 2013). Third, even if one were to take the critique seriously, there is arguably still no consensus on exactly how one should go about testing the quantity hypothesis. Michell's own proposals are largely based on testing axioms of additive conjoint measurement (see, e.g., Michell, 1990), but examples of the successful use of such approaches are still largely absent from the literature (for commentary about why this is the case, see, e.g., Borsboom, 2005, 2006; Cliff, 1992). Some scholars have proposed that the Rasch model can be interpreted as a probabilistic instantiation of the theory of additive conjoint measurement (e.g., Kyngdon, 2008; Perline et al., 1979) and as such can at least indirectly test the quantity hypothesis (Briggs, 2013; Domingue, 2014; Maul et al., 2019), though there remains controversy both regarding whether this is a reasonable proposal in principle (Michell, 2004) and to what extent the Rasch model is actually used this way in practice (e.g., McGrane & Maul, 2020). Other scholars have proposed alternative paths to establishing the meaningfulness of quantitative measurement results, such as Sherry (2011), who argued that such claims can be evaluated on pragmatic grounds by investigating their usefulness over weaker (i.e., nonquantitative) claims, though again, clear applications of such approaches in the human sciences seem to remain rare.

Returning to the concept of validity, Borsboom and Michell's perspectives (along with others not discussed at length here; e.g., Hughes, 2018; Maraun, 1996, 2012; Sijtsma, 2012; Trendler, 2009) could be seen as not necessarily incompatible with the argument-based perspective, but as focused on the precise nature of the evidence and argumentation that would be necessary to defend one particular type of interpretation of test scores. That is, it could be noted that interpreting test scores as measures of a construct is only one among many possible interpretations, and tests are routinely put to uses that, strictly speaking, do not appear to require that any measurement take place at all (as when, for example, the mere existence of a test inspires a student to study). That said, interpretations of test scores as measurement results do seem to be quite common, insofar as test scores or assessment results are often presented and interpreted in quantitative terms, and the perspectives of Borsboom and Michell call attention to the need to take these claims as being in need of rigorous evaluation in their own right.

From our perspective, this is where the contribution of the present chapter comes in. The TISM framework aims to help clarify the elements that secure the objectivity and trustworthiness of measurement results—that is, their validity, by another name. Again, we see this not as incompatible with the argument-based perspective on validation, but as complementary: If the argument for the adequacy of a test for a given purpose involves a claim about measurement, careful attention to and documentation of each of the TISM components and their interplay can provide backing for that argument. Conversely, failing to attend to all four elements of the framework could result in an argument that is incomplete at best or deceptive at worst.

Reliability, Measurement Error, and Uncertainty

If there is one procedure that any psychometrician working with test data knows how to complete, it is how to compute a *reliability coefficient* (usually Cronbach's alpha) and convert it into a standard error of measurement (SEM). In the context of tests that have been modeled with IRT, conditional SEMs can be examined through the inversion of a test information function. The SEMs associated with educational test scores are usually substantial, even when reliability coefficients are near .90. For example, on the Smarter Balanced ELA/Literacy tests taken by students in California, the marginal reliability coefficients for each grade are between .86 and .88, yet the average SEM ranges from a low of 35.0 in Grade 3 to a high of 36.6 in Grade 8. To put this magnitude in perspective, recall the mock Smarter Balanced ELA/Literacy score report for the Grade 5 student depicted in Figure 21.3. The hypothetical student on that report has a scale score value of 2421. Assuming that measurement error is normally distributed, if a single SEM of 36 is added and subtracted, the range of possible scores ranges from 2385 to 2457, which encompasses two of the four possible achievement levels (Levels 1 and 2); if two SEMs are added and subtracted, the range would be 2349 to 2493, encompassing three of the four possible levels (Levels 1, 2, and 3). Notably, representation of this uncertainty is missing from the mock score report in Figure 21.3. Although “Sophia’s” test score places her in Level 2 (“Standard nearly met”), she may well have ELA/Literacy knowledge and skills consistent with Level 3 (“Standard met”) or Level 1 (“Standard not met”). In short, there is considerable uncertainty associated with Sophia’s measured value of ELA/Literacy ability, and this uncertainty seems to be substantially larger than is typically associated with physical measuring instruments such as thermometers.

In metrology, the concept of uncertainty is somewhat more encompassing compared to what is referred to as *measurement error* in psychometric contexts. In metrology, it is common to see total “uncertainty budgets” composed of several distinct forms of uncertainty, such as uncertainty associated with the way in which the property is defined (*definitional uncertainty*), the physical limitations of the measuring instrument (*instrumental uncertainty*), and the ways in which scales are defined with respect to a common unit or standard (*calibration uncertainty*; see, e.g., Mari et al., 2023). Of these, what psychometricians refer to as measurement error most closely resembles instrumental

uncertainty; even more specifically, instrumental *repeatability* is associated with the independent replication of the same measurement procedure across occasions. In educational measurement, where independent replications across occasions are difficult to realize, the concept is often understood with respect to variation in the items that have been selected to comprise the test instrument. Disentangling uncertainty due to other facets associated with an instrument's repeatability (specific to, for example, choices of test occasion, mode of presentation, linking items), would require a more complex experimental test design and mathematical model (for examples, see the many-faceted Rasch model, Linacre, 1989; generalizability theory, Brennan, 2003; or the variance decomposition approaches used for NAEP, Oranje & Kolstad, 2019).

The TISM framework calls attention to the fact that there are several forms of potential uncertainty associated with any given set of measurement results, and some of these might be difficult to quantify. With respect to theory, there might be substantial *definitional uncertainty*: unlike temperature, which is understood to be an invariant feature of the world and of which we now have a fairly stable and detailed understanding, both our conception of and the actual nature of reading comprehension continue to evolve. Even if we were to limit ourselves to a definition of reading ability as primarily a function of text complexity, there is heated debate among reading experts about the proper way to define text complexity (cf., Graesser et al., 2004; Sheehan, 2013; Stenner et al., 1983). If different definitions of text complexity chosen by test developers lead to different choices in the construction of reading passages and test questions, it may not be obvious how or even whether results can be compared across different tests, which is a well-known problem with operationalism in general (see, e.g., Bickhard, 2001; Borsboom, 2005, chap. 2; Markus & Borsboom, 2013, chap. 2; Mari et al., 2023, chap. 4). Even if the measurement results could be considered comparable in principle, these choices could change the ordering of item difficulty, which could change the location of achievement-level thresholds as part of a standard-setting procedure.

There is also ample room for disagreement when explicating a broader sociocultural theory of reading comprehension: Does it include reading speed? Are orthographic fluency and morphemic awareness subcomponents of reading comprehension or causes of it? Does reading comprehension include metacognitive abilities like critical analysis of textual information, or should the latter be regarded as a distinct but related construct? And so on. It is rare to find that a testing program has taken a clear stance on these substantive issues. In short, the theoretical and definitional uncertainty in attempting to characterize a construct such as reading comprehension can be considerable, and because theories and definitions of educational constructs are rarely specified with respect to an explicit mathematical model, it will usually only be possible to express this uncertainty in qualitative terms. As a result, what is referred to as measurement error in psychometric contexts may relate to only a small portion of the total uncertainty that should be attributed to results of an educational measurement.

Scales, Units, and the Evaluation of Magnitude

One of the most fundamental differences between measuring scales in the physical sciences and the “score scales” on which test performance is expressed in educational contexts is in the interpretation that can be afforded to the units of measurement. In the physical sciences, although the specific choice for a unit is sometimes arbitrary (e.g., Celsius vs. Fahrenheit), the ability to connect that unit to theory such that it can be realized experimentally is not, and instruments are carefully calibrated to ensure that the meaning of a unit is the same across different replications of the same instrument. Most well-known measuring scales in the physical sciences have ratio or interval properties. The properties of educational measurement scales are far more uncertain. In general, they are presumed to have at least ordinal properties, but are frequently interpreted and used in statistical analyses in ways that imply interval properties. To illustrate this, Figure 21.5 provides an example of how the results from the 2022 NAEP Reading Assessment were publicly disseminated for Grade 4 students.

The graphic in Figure 21.5 emphasizes the comparison of the average scale score to three other time points, most notably 2019 (a year before the onset of the COVID-19 pandemic), 2003 (about 20 years earlier), and 1992 (the first score available in the trend, about 30 years earlier). Given the widespread disruptions to schooling and learning experiences between 2020 and the time during which students were tested for NAEP in January 2022, it is not surprising to see that average scores have declined.

But how should we evaluate the *magnitude* of a 3-point decline? In Figure 21.5, it is only apparent that a 3-point decline is “big” by comparing it to all other score changes over a 2- to 4-year interval. Without seeing this, how should we evaluate the magnitude of 3 scale score units? One way to attempt to answer this question would rely on achievement-level thresholds (i.e., *cut scores*) on the NAEP score scale to serve as de facto reference standards. Specifically, a score of 208 is the threshold for achievement

Trend in fourth-grade reading average scores

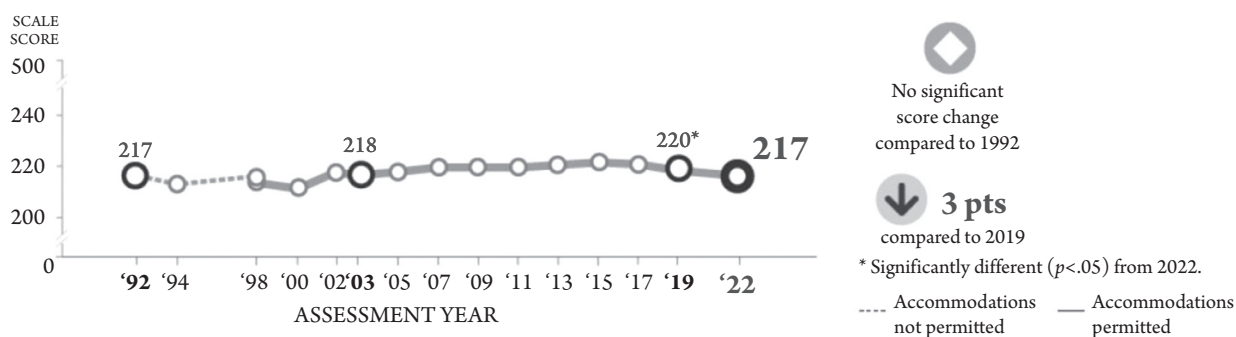


FIGURE 21.5
Public Dissemination of Results From National Assessment of Educational Progress 2022
Reading Assessment

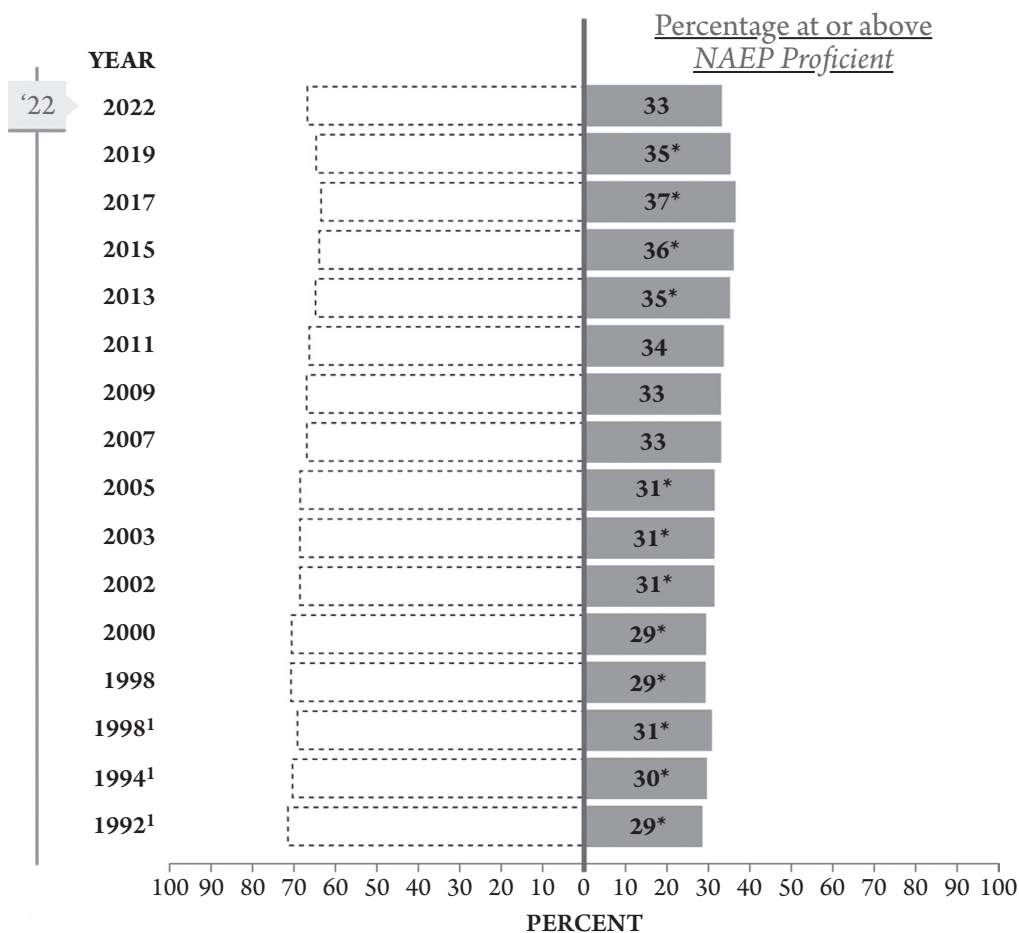
Note. Data are from https://www.nationsreportcard.gov/mathematics/supportive_files/2022_rm_infographic.pdf, retrieved July 16, 2023.

that is considered *Basic* (as opposed to *Below Basic*), a score of 238 is the threshold for achievement that is considered *Proficient* (as opposed to *Basic*), and a score of 268 is the threshold for achievement that is considered *Advanced* (as opposed to *Proficient*). From this, we can infer that the national average for reading comprehension has landed between the demarcations for the *Basic* and *Proficient* achievement levels since 1992. We can also notice that the distance between the *Basic* and *Proficient* demarcations is 30 points, which puts into perspective that a change of 3 points is 10% of the distance between two particular locations along the score scale that have been designated as qualitatively meaningful. Interestingly, the distance between the *Proficient* and *Advanced* thresholds happens to also be 30 points. This gives the impression that the difference in reading comprehension necessary to move from a *Basic* to a *Proficient* level is equivalent to the amount needed to move from a *Proficient* to an *Advanced* level.

Figure 21.6 contrasts with Figure 21.5 in that it depicts the trend in reading ability only with respect to the (inferred) percentage of fourth-grade students with a score above the demarcation for the *Proficient* achievement level. In effect, each student's test score is reduced to a binary variable that takes a value of 1 if the score is 238 or higher and a value of 0 otherwise. From this, we can see that a 3-point decrease on the NAEP scale from 2019 to 2022 translates into a 2 percentage point decrease in the number of students with test performance that would place them in one of the top two NAEP achievement levels.

Now, strictly speaking, if the NAEP scale has interval properties, both Figures 21.5 and 21.6 would represent equally defensible ways to convey information about educational progress (or lack thereof) over time; the choice between them simply depends on what information one wishes to emphasize. This is because one can always choose to “go backward” from a scale that falls along Stevens's taxonomy (turning an interval scale into an ordinal scale). However, if the NAEP scale only has ordinal properties, then a representation such as the one shown in Figure 21.6 might be more defensible because it places a focus on changes in a categorical frequency distribution instead of a change in means. When a scale is ordinal, then any monotonic (i.e., order-preserving) transformation to it is admissible, and depending on the nature of the transformation, a comparison between two individuals or groups on the untransformed and transformed scales using means and variance can lead to contradictory conclusions (see, for example, Michell, 1986). Therefore, if the NAEP scale is only ordinal, it could be transformed in a way that led to the conclusion that the mean scores had decreased by more than 3 points during the pandemic—or that the mean had stayed the same, or even increased—without any loss of information about order. This means that the answers to questions about magnitude (i.e., practical significance) that are based on a comparison between mean scores on an ordinal scale lack generalizability, and this is why questions about scale properties are fundamental to scientific research in education.

Although the question of whether a scale has interval as opposed to ordinal properties is important, a superficial focus on this question can obscure important theoretical

**FIGURE 21.6**

Public Dissemination of National Assessment of Educational Progress Reading Results in the Metric of Percentage Proficient or Advanced

Note. Data are from <https://www.nationsreportcard.gov/reading/nation/achievement/?grade=4>, retrieved July 16, 2023.

*Significantly different ($p < .05$) from 2022.

¹Accommodations not permitted.

assumptions being made about the construct of measurement. That is, to focus solely on methods that can be used to generate a numeric scale (i.e., the I, S, M ellipses of the TISM framework) is to ignore the interactions of these methods with substantive theory. For example, in a discussion about NAEP scale score properties, Zwick (1992, p. 209) wrote that “when the characteristic to be measured cannot be directly observed, claims of equal-interval properties with respect to that characteristic are not testable and are therefore meaningless” and, further, that the “interval property has no particular implications for the relation between the units of the scale and the attributes the scale purports to measure or predict”—in other words, claims about interval properties are untestable, meaningless, and without consequence. Now, if the construct of measurement is hypothesized to be a quantity, this is a curious position to take. After all, if a scale

is claimed to have interval properties, there are clear implications for the relationship between differences in the construct and differences on the scale. For instance, if *Basic*, *Proficient*, and *Advanced* represent qualitatively ordered levels of a reading comprehension construct, then if an interval scale indicates that *Proficient* is 30 units from *Basic*, and *Advanced* is 30 units from *Proficient*, this would mean that the *difference* in reading comprehension between these levels is the same. One source of confusion here is with the idea that the interval properties of a scale can only be established when a one-to-one correspondence can be “directly observed” between the units of the measuring scale and a decomposition or recomposition of the construct of interest. Yet there is no way to directly compare the intervals on a Celsius thermometer to changes in temperature, nor is it possible to directly compare the seconds on a watch to changes in time. Still, we accept these instruments as producing measures of quantitative properties on an interval or ratio scale because of a theoretical understanding about the relationship between thermodynamic energy, liquids, and volume in the case of temperature and because we have engineered an invariant meaning for the second from the vibrations of a cesium atom. Moreover, there is a substantial body of literature on how scale properties can be (indirectly) tested, as briefly discussed in our section on the Rasch Model. It therefore goes too far to suggest that a construct such as reading comprehension cannot possibly be expressed on an interval scale because the latter is not “directly observable.”

However, the conclusion that forming an interval scale for an educational construct is *possible* does not imply that it is *plausible* (e.g., see Michell, 2014). In educational contexts there is seldom any formal hypothesis in place about the structure of a construct like reading comprehension. Even when learning theories are most consistent with a model of the construct that posits ordinal levels with heterogeneous differences in terms of the cognitive processes involved, the construct posited as underlying the responses to an educational test is typically treated “as if” it was a continuous quantity (Lord & Novick, 1968; Mislevy, 2018). This may be done to take advantage of the putative affordances from IRT models for activities such as adaptive test design and equating test forms composed of different sets of items (the I-by-M interaction of the TISM framework). But the potential incoherence of the model with substantive theory (the T-by-M interaction of the TISM framework) is one reason why reading scholars may be skeptical of claims that reading comprehension has been accurately measured by a single number of a reporting scale.

In much of the educational scaling literature, either the issue of how one would distinguish ordinal from interval scales is not taken up at all (e.g., Kolen & Brennan, 2004) or the advice provided seems fairly ambiguous. For example, Lord and Novick (1968, pp. 21–22) suggested that any test score scale can be treated as if it were on an interval scale so long as it can be shown to be a good predictor of some behavioral criterion. But this simply pushes the question back to the choice of behavioral criterion and the evidence needed to establish “goodness” of prediction. If the prediction involved were to be evaluated based on, say, the value of a Pearson correlation coefficient, this would appear to presume from the outset that the test and criterion variables are on at least interval scales. Moreover, the interpretation of the result is equivocal: Should a weak correlation

be taken as evidence that the scale of the test score (and/or the criterion score) does not have interval properties? Or should it be taken as evidence of invalidity—that the test does not have the expected relationship with a criterion it was expected to predict? A problem with the pragmatic advice from Lord and Novick is that any substantive theory one might have about the construct does not appear to have any falsifiable implications for the scale and unit in which measurements are to be reported.

Another matter of debate in the field has to do with the relationship between desired scale properties and the choice of mathematical model applied to test responses (the S-by-M interaction of the TISM framework). In one tradition, consistent with Lord and Novick's approach, choice of scale is considered independent of model selection (for details, see Briggs, 2021a). This is not the case in the Rasch measurement tradition, where adequate fit to the Rasch model is taken as a necessary condition for a scale with interval properties.

To the extent that IRT can serve as a tool for establishing a measuring scale with a designated and invariant reference interval, the Rasch model does in fact come with clear advantages relative to models such as the 2PLM and 3PLM (used for Smarter Balanced and NAEP tests, respectively). Under the Rasch model, the comparison of any two individuals is invariant to the choice of item that serves as a frame of reference, and by extension, the comparison is invariant to the choice of any distance between two items. When applying the Rasch model to construct a scale, any distance between two items (or between groups of items) can be used as an interval to compare distances between individuals.¹⁰ But this is contingent on having items on a test instrument that maintain the same order of response probabilities at any location on the scale. In short, the item characteristic curves cannot cross. If they do, as can be the case under the 2PLM and 3PLM, then the interpretation of differences between individuals will be sensitive to the choice of a reference interval, and it will be challenging, if not impossible, to articulate the meaning of a unit across the full range of the resulting scale. This is not to say that the Rasch model can perform alchemy. It imposes strong conditions on item design that can be extremely challenging to satisfy in practice, especially in the absence of a theory that can be used to manipulate factors that predict item difficulty. In the absence of such theory, fitting the Rasch model in isolation will be insufficient to establish an *interpretable* and *replicable* unit of measurement (see, e.g., Briggs, 2019).

One example of an approach that relies on a theoretically motivated application of the Rasch model is the Lexile scale for reading comprehension (Stenner et al., 2013; Stenner et al., 1983). The Lexile scale was explicitly developed to be analogous in at least some respects to the measurement of temperature. As described previously, the intervals for the Celsius scale are identified relative to anchor points (i.e., reference standards) that correspond to discernible thermodynamic phenomena associated with the freezing and boiling points of water, with a unit then defined relative to these standards (e.g., a degree Celsius is 1/100th of the distance between these two reference standards). The Lexile scale attempts to mimic this approach first by modeling reading ability and item difficulty in terms of two central factors that make texts more or less complex: average sentence length in a reading passage and average familiarity of the

words in the passage. These factors are used to connect the difficulty of items calibrated to a Rasch model directly to the text complexity of reading materials associated with different levels of schooling. A Lexile unit is then defined as 1/1,000th of the distance between the text complexity a student should be expected to comprehend in the first grade of schooling and the text complexity of passages found in an electronic encyclopedia, typical of the midpoint of reading materials found in workplace text (also typical of text found in grade 10 reading passages).

Note that the two reference locations on the Lexile scale are more equivocal than the reference locations for a thermometer because they depend on societal judgments about what children should be able to read at different ages, and this is likely to change over time. Beyond this, the extent to which the instrumentation used to establish the Lexile scale fits the Rasch model and/or adequately satisfies the axioms of additive conjoint measurement remains an open question (cf., Domingue, 2014; Kyngdon, 2011). But the general approach in which a scale and its units are being motivated by an interaction between theory and modeling (a T-by-S-by-M interaction in the TISM framework) represents a valuable proof of concept.

As we have seen, both the Smarter Balanced and the NAEP scales also have reference standards that have been defined through the process of writing achievement-level descriptors in concert with item analysis. However, these demarcations have no bearing on the way the unit of each measuring scale is defined. For example, the unit of the 0 to 500 NAEP score scale was originally chosen so that a value within this range could be interpreted as the number of items the average student would answer correctly on a hypothetical test with 500 items (Beaton & Johnson, 1992). The unit of the Smarter Balanced ELA/Literacy scale was defined by transforming the logit estimates from a calibration of the underlying item response model such that a change of one base unit (1 logit) is the equivalent of a change on the operational scale of 79.3 points. However, no rationale for this transformation is available in published materials, nor is any attempt made to provide an interpretation for the new unit that this transformation implies.

One promising approach that has been used to enhance the interpretability of the NAEP reporting scale is known as item mapping,¹¹ in which exemplar items are used to illustrate the kinds of questions that students would be expected to answer correctly at a given location of the scale. As implemented for NAEP, this approach makes no commitment to the scale having interval or even ordinal properties, and indeed because the locations of items on the scale correspond to a fixed response probability for the 3PLM, if a different response probability were selected to characterize item difficulty, many of the items might change their order (another instance of an S-by-M interaction in the TISM framework).

The need to make inferences about growth and development has always been a primary motivation for educational measurement (Rasch, 1980; E. L. Thorndike, 1918). But without careful attention to scales and units, such inferences will be highly equivocal. For example, meta-analysis is a method invented to permit generalizations about

the magnitude of experimental treatments, which in education research could include a wide range of programs and curriculum intended to have positive effects on student learning. When test scores serve as outcome measures, they are ostensibly placed on a common scale by expressing each difference in means with respect to a test's standard deviation (i.e., *effect size units*). But if a collection of experimental studies come from tests that convey only ordinal information, a variety of admissible monotonic transformations to the scales could inflate or deflate these standard deviations, which would in turn influence effect size estimates and thereby distort conclusions about the average size of the effects inferred from the meta-analysis.

Almost all educational measurement procedures use instrumentation to convert observable information about order into a location on a measuring scale according to a designated unit or into a scale in which a categorization is made relative to one or more reference standards. In educational testing, the score given to a student's test item response provides information about order, and a full test provides information about patterns of order in some vector of item responses. What can be left unsaid is whether and how evidence of order, via some mathematical model, is intended to provide falsifiable implications about a measurement construct. Attention to the TISM framework can bring these sorts of issues and conceptual tensions to the surface.

LOOKING FORWARD

We hope that this chapter has provided an entry point for those who appreciate—or are coming to appreciate—the complexity of measurement as both a concept and a process. Educational measurement is not just a concept, but also an established field of research and practice. If the field is to evolve and thrive, we would argue that it is important for professionals within this community to continue to grapple with its core concepts. Outside a rather small niche of scientists and scholars, educational measurement is most commonly described in a manner that paraphrases Stevens's claim that measurement is a process of assigning numbers to objects according to a specified procedure. Historically and into the present day, considerable emphasis has been placed on the design and administration of educational tests (i.e., *instrumentation*) and the *statistical modeling* of the scores that result. Thus, when a student takes a test, it does indeed produce a numeric value that gets assigned back to the student or some aggregate of students. It is little wonder in this sense that testing and measurement are viewed as one and the same thing. We hope that after reading our chapter the reader can appreciate why this conflation of terms is problematic.

We take a more encompassing view of the process of measurement as one that

- provides information about a property of an object or event in the world;
- uses instrumentation designed to be sensitive to differences in this property; and

- produces information in the form of values of that construct related to a scale or unit, which is, in turn, qualified by information about uncertainty.

Each of the elements in this process has a direct connection with one of the four components of the TISM framework we have introduced in this chapter. In this framework, we locate the practice of measurement in the interacting bodies of knowledge and practice that pertain to substantive theory, instrumentation, scales and units, and modeling. Many prominent perspectives on measurement in the human sciences call attention to one or more of these components, but leave others in the background, which can lead to incomplete or erroneous justifications for measurement claims. By “zooming out” to a structural understanding of measurement like the one we propose here, we gain a greater understanding of and appreciation for what is at stake with respect to the measurement debates and the conditions that would need to be met for claims of measurement to be fully justified. In an educational context, tests have the potential to play a critical role as the instrumentation of a measurement process, but in the absence of interactions with substantive theory, scales and units, and models, a test is just a series of questions, and there is no compelling reason to dress it up with the prestige that the term measurement is often intended to confer.

It has been over 50 years since the chapter “The Nature of Measurement” by Lyle Jones appeared in the second edition of *Educational Measurement*. At this time, Jones defined measurement as

a purposive acquisition of information about the object, organism, or event measured by the person doing the measuring. It is a determination of the magnitude of a specified attribute of the object, organism, or event in terms of a unit of measurement. The result of the measurement is expressed by a numeral. The classification of attributes, either qualitative or quantitative, is distinguished from the measurement of attributes, which must be quantitative. (Jones, 1971, pp. 336–337)

The reader will notice that Jones explicitly distinguished measurement from classification. In this chapter, we have not taken a firm stance on what is and is not an instance of measurement or, for example, whether the administration of a test for the purpose of classifying subjects according to discrete attributes (i.e., data that are typically modeled using latent class methods) is justifiably characterized as measurement. One reason for this is that although we agree in the broad sense expressed by the TISM framework on the necessary conditions for measurement, we would each likely give different answers about the sufficient ones. Another reason is that a broadened conceptualization of measurement, one generally in accord with Stevens’s definition, has long since become entrenched in public and professional discourse. Even within metrology, a field devoted to the scientific study of measurement and its applications in the physical sciences, an orderable property is also referred to as a quantity, and it is considered sensible to speak of measuring it (see Joint Committee for Guides in Metrology, 2012, pp. 2–3). There is

little point in attempting to police the usage of measurement terminology as it is applied to educational testing. The cat is out of the bag.

There is, however, great value in the recognition that measurement is a complex endeavor that goes beyond clever instrumentation and sophisticated statistical models. Although the terms *test*, *assessment*, and *measurement* may be used synonymously in public discourse, it is important for a professional community involved in educational measurement (e.g., NCME) to be able to articulate the salient differences. A test is just a set of questions that are subject to evaluation. An assessment is the evaluation or estimation of the nature, quality, or ability of someone or something. All measurement requires assessment, but not all assessment requires measurement. In classroom and informal educational settings, assessments are made all the time; in our view, it takes a more concerted effort to convert an assessment into a measurement.

Measurement has earned its social capital largely because of its highly successful history in the physical sciences and engineering. When educators and other human scientists make claims to measurement, they are at least implicitly laying claim to this capital, in particular, to the claim that measurement produces trustworthy information about a measured construct in the form of values of that construct. However, given the many differences between measurement in the physical and human sciences, it is not obvious that this trustworthiness has the same warrant. Indeed, an exploration of the methods and activities that have been described as measurement in educational and psychological contexts would be incomplete without an acknowledgment of the way that premature and ill-founded claims related to the measurement of intelligence provided a scientific veneer to racist, sexist, and xenophobic policies and practices in the early 20th century (see Briggs, 2021b; Gould, 1981; Stein, 2016). A lesson from this historical episode comes in the danger of reifying the name given to a test, that is, conflating it with the construct (or constructs) the test was designed to measure. More generally, when it comes to measurement, a great deal of humility is in order, especially when a theoretical understanding of the construct is nascent or lacking altogether. It is worth reiterating that although the evolution of our ability to measure temperature represents an accomplishment worthy of emulation, it is one that required hundreds of years of study and development. Something analogous in education may seem unlikely for many reasons, an obvious one being that unlike thermal energy, the constructs at play are socially constructed and likely to change over time. But without an ongoing commitment to a theoretical understanding of constructs we choose to value and the methods we employ to measure them, measurement will never be more than an operational activity. We argue it is better to appreciate measurement as an aspirational activity, one that is cyclical and iterative, even if (or when) it is being undertaken for pragmatic rather than scientific purposes.

A limitation to the TISM framework we have introduced here is that it may give the impression that measurement procedures can be isolated from social, cultural, historical, and political contexts. This is clearly not the case. Consider the notion of objectivity, which has long been taken to be a hallmark of scientific measurement. Even when measurement instrumentation may appear to be objective, whether in the physical sciences

or in education, the choices of what to measure, who to measure, and the purpose for measuring are often highly subjective and value laden, and how this is negotiated can be political and contentious.

A recent example can be found in the process that culminated in a revision to the NAEP Framework for Reading in 2021. All subject domains tested on NAEP come with a framework that serves as the basis for item design. The process of revising a framework is overseen by NAGB, which consists of 26 members who are appointed by the U.S. Secretary of Education to set policy for NAEP. An initial framework is drafted by the members of a development panel, convened through nomination to be representative of subject matter experts, practitioners, administrators, researchers, business representatives, and members of the general public. This draft framework is then made available for comment and feedback from both NAGB and the general public. The feedback is used to revise the framework, and a final version must be approved by NAGB before it can go into effect.

The original version of the reading framework that was submitted for public comment foregrounded a concern about equity in education and the importance of design following from a sociocultural theory of learning as two new emphases for the 2026 NAEP Reading tests. These foci, among other features of the draft framework, evoked a controversial reception during public commentary. Questions were raised about what it meant for a measuring instrument to be “equitable by design” and whether this was consistent with the purpose of NAEP. Public officials from conservative states took issue with the term *sociocultural*, worrying that NAEP was becoming a vehicle for progressive social values. This influenced subsequent revision of the framework. Table 21.1 compares the definitions of reading comprehension provided in the original and revised reports.

Table 21.1 The Politics of the National Assessment of Educational Progress (NAEP) Reading Framework Revision	
Definition of Reading Comprehension Submitted to NAGB by Reading Specialists on the NAEP Framework Development Panel (June 2020)	Definition of Reading Comprehension Approved by NAGB (September 2021)
Reading comprehension is a sociocultural process in which individuals use language, knowledge, and foundational skills to extract, construct, integrate, and critique meaning as they engage with a wide range of texts for purposes shaped by home, community, and school experiences.	Reading comprehension is making meaning with text, a complex process shaped by many factors, including readers’ abilities to: • engage with text in print and multimodal forms; • employ personal resources that include foundational reading skills, language, knowledge, and motivations; and • extract, construct, integrate, critique, and apply meaning in activities across a range of social and cultural contexts.

To be clear, the definition of reading comprehension that was ultimately approved by NAGB is not inconsistent with a sociocultural theory of reading, but there are subtle and substantive differences in wording that reading and literacy scholars would notice, even if the lay public would not. In short, the substantive theory that is made publicly visible as a basis for test design on NAEP (the T of our TISM framework) is one that is mediated by both values and political power in ways that can be difficult to predict. There are similar stories to be told any time a test is administered on behalf of a public entity to monitor and evaluate trends in educational achievement.

What does the future hold for educational measurement? In the first half of the 20th century, the technological innovation of the group-administered standardized test gave birth to an educational testing industry, and modeling innovations followed to solve the problems that arose in the practice of testing students on an unprecedented scale. In the first half of the 21st century, we are in the midst of new technological innovations that could have dramatic ramifications for educational measurement: computers, the Internet, social media, and, most recently, artificial intelligence. As individuals increasingly experience assessment activities in digital environments, there will be both more and different kinds of data amenable to empirical analysis (e.g., big data and computational psychometrics; see, for example, von Davier et al., 2021). Whether this will lead to improvements in educational measurement is another question (see Briggs, 2022). The ability to make trustworthy inferences on the basis of test results depends on the quality of stimuli that comprise the test. More and different kinds of data may not provide any novel insights about student learning, especially when they are not integrated coherently with theories of learning. Will big data, digital learning environments, and artificial intelligence lead to better educational measurement? This will be an interesting and open question. What we can be most confident about is that there will always be a need to evaluate the methods, practices, and approaches to teaching and learning. It is hard to imagine a world in which a comprehensive evaluation could be done without an attempt at educational measurement. To do this will continue to require a reckoning with the four central components of measurement: theory, instrumentation, scales and units, and modeling.

ACKNOWLEDGMENTS

We would like to thank Robert Brennan, Neil Dorans, Mark Wilson, and the editors for the constructive comments provided on preceding drafts of this chapter.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.
- Ackerman, T.A., Bandalos, D. L., Briggs, D. C., Everson, H. T., Ho, A. D., Lottridge, S. M., Madison, M. J., Sinharay, S., Rodriguez, M. C., Russell, M., von Davier, A. A. & Wind,

- S. A. (2024). Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice*, 43, 7–17. <https://doi.org/10.1111/emip.12581>
- Alexander, P., & Fox, E. (2004). A historical perspective on reading research and practice. In R. Ruddell & N. Unrau (Eds.), *Theoretical models and processes of reading* (5th ed., pp. 33–68). International Reading Association.
- Andrich, D. (1988). *Rasch models for measurement*. Sage.
- Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1), II-7–II-16. <https://doi.org/10.1097/01.mlr.0000103528.48582.45>
- Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer. <https://doi.org/10.1007/978-981-13-7496-8>
- Angoff, W. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 506–600). American Council on Education.
- Baker, F. B. (1985). *The basics of item response theory*. Heinemann.
- Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed). Marcel Dekker.
- Beaton, A. E., & Johnson, E. G. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, 29(2), 163–175.
- Bich, W., Cox, M. G., & Harris, P. M. (2006). Evolution of the “Guide to the Expression of Uncertainty in Measurement.” *Metrologia*, 43(4), S161.
- Bickhard, M. H. (2001). The tragedy of operationalism. *Theory and Psychology*, 11(1), 35–44.
- Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison–Wesley.
- Bock, R. D. (1997). A brief history of item theory response. *Educational Measurement: Issues and Practice*, 16(4), 21–33.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4), 443–459. <https://doi.org/10.1007/BF02293801>
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge, Taylor & Francis Group.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440.
- Borsboom, D. (2009). *Educational Measurement*, 4th edition [Book review]. *Structural Equation Modeling*, 16, 702–711.
- Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10(1–2), 38–41. <https://doi.org/10.1080/15366367.2012.681971>

- Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 135–170). Information Age Publishing.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38, 295–317.
- Brennan, R. L. (2003). *Generalizability theory*. New York, NY: Springer.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50, 204–226.
- Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch model. *Measurement*, 146, 961–971.
- Briggs, D. C. (2021a). A history of scaling and its relationship to measurement. In B. Clauser & M. Bunch (Eds.), *The history of educational measurement* (pp. 263–291). Routledge.
- Briggs, D. C. (2021b). *Historical and conceptual foundations of measurement in the human sciences: Credos and controversies*. Routledge.
- Briggs, D. C. (2022). NCME Presidential Address 2022: Turning the page to the next chapter of educational measurement. *Journal of Educational Measurement*, 59(4), 398–417.
- Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, 38(6), 551–576.
- Brogden, J. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, 42(4), 631–634. <https://doi.org/10.1007/BF02295985>
- Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, 75(1), 33–57.
- Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307–335.
- Cai, L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring* [Computer software].
- Campbell, N. R. (1920). *Physics: The elements*. Cambridge University Press.
- Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. Longman, Green & Co.
- Campbell, N. R. (1938). Symposium: Measurement and its importance for philosophy—Introduction. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 17, 121–141.
- Cattell, J. M. (1890). Mental tests and measurements. *Mind*, 15, 373–381.
- Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
- Chang H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1), 1–20. <https://doi.org/10.1007/s11336-014-9401-5>

- Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage Publications.
- Clauser, B. (2021). A history of classical test theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 157–180). Routledge.
- Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3(3), 186–190.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). American Council on Education.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. Wiley.
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). American Council on Education.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Science & Business Media.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1), 1–38. <http://www.jstor.org/stable/2984875>
- Domingue, B. (2014). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*, 79(1), 1–19.
- Embretson, S. E. (Ed.). (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.
- Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38(4), 343–368.
- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.
- Euclid's *Elements of geometry*, the Greek text of J. L. Heiberg (1883–1885) edited, and provided with a modern English translation, by Richard Fitzpatrick, 2008. <https://farside.ph.utexas.edu/Books/Euclid/Elements.pdf>
- Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houstoun, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H., & Tucker, W. S. (1938). Quantitative estimates of sensory events: Interim report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *British Association for the Advancement of Science*, 108, 277–334.
- Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., & Tucker, W. S. (1940). Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Report of the British Association for the Advancement of Science*, 2, 331–349.
- Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, 7, 19–29.

- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. Springer.
- Galton, F. (1875). Statistics by intercomparison with remarks on the law of frequency of error. *Philosophical Magazine*, 49, 33–46.
- Galton, F. (1877). Address to the Anthropological Department of the British Association. William Clowes and Sons.
- Galton, F. (1885). On the anthropometric laboratory at the Late International Health Exhibition. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 14, 205–221.
- Goddard, H. (1910). Four hundred feeble-minded children classified by the Binet method. *Pedagogical Seminary*, 17, 387–397.
- Godfrey-Smith, P. (2003) *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25(4), 21–35.
- Gould, S. J. (1981). *The mismeasure of man*. Norton.
- Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights From the Behavioral and Brain Sciences*, 2, 42–50. <https://doi.org/10.1177/2372732215600888>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36, 193–202. <https://doi.org/10.3758/BF03195564>
- Guilford, J. P. (1936). *Psychometric methods*. McGraw–Hill.
- Gulliksen, H. (1950). *Theory of mental tests*. Routledge.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Greenwood Publishing Group.
- Holland, P. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577–601.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Greenwood Publishing Group.
- Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale, and test development* (pp. 751–779). John Wiley & Sons.
- Humphry, S. (2011). The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research and Perspectives*, 9(1), 1–24.
- Humphry, S. (2013). A middle path between abandoning measurement and measurement theory, *Theory & Psychology*, 23(6), 770–785.

- Joint Committee for Guides in Metrology. (2012). *JCGM 200:2012, International vocabulary of metrology—basic and general concepts and associated terms (VIM)* (3rd ed.). (2008 version with minor corrections). <http://www.bipm.org/en/publications/guides/vim.html>
- Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike & W. H. Angoff (Eds.), *Educational measurement* (2nd ed., pp. 335–355). American Council on Education.
- Jones, L. V., & Thissen, D. (2006). A history and overview of psychometrics. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1–27). Elsevier.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Greenwood Publishing Group.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73. <https://doi.org/10.1111/jedm.12000>
- Kelley, T. L. (1923). The principles and technique of mental measurement. *American Journal of Psychology*, 34(3), 408–432.
- Kelley, T. L. (1927). The principles and technique of mental measurement. *American Journal of Psychology*, 34(3), 408–432.
- Kendeou, P. (2020). The assessment of reading for understanding. In P. D. Pearson, A. Palincsar, S. G. Biancarosa, & A. Berman (Eds.), *Reaping the rewards of the reading for understanding initiative* (pp. 67–142). National Academy of Education.
- Kevles, D. J. (1968). Testing the army's intelligence: Psychologists and the military in World War I. *Journal of American History*, 55, 565–581.
- Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85(5), 363–394.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Greenwood Publishing Group.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Springer.
- Krantz, D. H., Suppes, P., Luce, R. D., & Tversky, A. (1971). *Foundations of measurement*, Vol. 1. Academic Press.
- Kridel, C. (2000). "Rice, Joseph Mayer." In *American national biography online*. Oxford University Press. <http://www.anb.org/articles/09/09-01039.html>
- Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, 18, 89–109.
- Kyngdon, A. (2011). Plausible measurement analogies to some psychometric models of test performance. *British Journal of Mathematical and Statistical Psychology*, 64(3), 478–497.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511815355>

- Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 61, 273–287.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice*, 39, 8–21. <https://doi.org/10.1111/emip.12318>
- Linacre, J. M. (1989). *Many-faceted Rasch measurement* [Doctoral dissertation, The University of Chicago].
- Lindquist, E. F. (1951). *Educational measurement*. American Council on Education.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, No. 7. (pp. 84).
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison–Wesley.
- Lorge, I. (1951). The fundamental nature of measurement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 533–559). American Council on Education.
- Luce, R. D., & Suppes, P. (2002). Representational measurement theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (pp. 1–41). John Wiley & Sons. <https://doi.org/10.1002/0471214426.pas0401>
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research*, 31(4), 603–616.
- Maraun, M. D. (2012). Validity and measurement. *Measurement: Interdisciplinary Issues and Perspectives*, 10, 80–83.
- Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy of Science*, 65, 46–56.
- Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement* (2nd ed.). Springer.
- Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.
- Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, 23(6), 752–769. <https://doi.org/10.1177/0959354313506273>
- Maul, A., Mari, L., & Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement*, 131, 764–770. DOI: 10.1016/J.Measurement.2018.08.068
- Maul, A., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement*, 79, 311–320.
- Maxwell, J. C. (1873). *A treatise on electricity and magnetism*. Clarendon Press.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

- McGrane, J. (2015). Stevens' forgotten crossroads: The divergent measurement traditions in the physical and psychological sciences from the mid-twentieth century. *Frontiers in Psychology*, 6, 431.
- McGrane, J., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, 152, 107346. <https://doi.org/10.1016/j.measurement.2019.107346>
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194. <https://doi.org/10.1177/01466219922031310>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100, 398–407.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Psychology Press.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88, 355–383.
- Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge University Press.
- Michell, J. (2005). The logic of measurement: A realist overview. *Measurement*, 38(4), 285–294.
- Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspective*, 6(1), 7–24.
- Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 111–133). IAP Information Age Publishing.
- Michell, J. (2014). The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma. *Theory & Psychology*, 24(1), 111–123. <https://doi.org/10.1177/0959354313517524>
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49(3), 359–381. <https://doi.org/10.1007/BF023>
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131–154. <https://doi.org/10.2307/1165166>
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
- Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 47(4), 355.
- Mosier, C. I. (1941). Psychophysics and mental test theory: II. The constant process. *Psychological Review*, 48(3), 235–249.

- National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. The National Academies Press.
- National Assessment Governing Board. (2021). *Reading framework for the 2026 National Assessment of Educational Progress*.
- National Research Council. (2000). *How people learn: Brain, mind, experience, and school* (Expanded ed.). The National Academies Press.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press.
- Newton, P. E., & Shaw, S. D. (2012). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–18.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1), 1–13. <https://doi.org/10.1007/BF02289400>
- Oranje, A., & Kolstad, A. (2019). Research on psychometric modeling, analysis, and reporting of the National Assessment of Educational Progress. *Journal of Educational and Behavioral Statistics*, 44(6), 648–670. <https://doi.org/10.3102/1076998619867105>
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22–37. <https://doi.org/10.1080/10888438.2013.827687>
- Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3(2), 237–255.
- RAND Reading Study Group. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. RAND Corporation.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (Expanded edition). University of Chicago Press.
- Rice, J. M. (1913). *Scientific management in education*. Publishers Printing Company.
- Richardson, M. W. (1936). Notes on the rationale of item analysis. *Psychometrika*, 1(1), 69–76.
- Rosenberg, A., & McIntyre, L. (2020). *Philosophy of science: A contemporary introduction*. Routledge.
- Samejima, F. (2010). The general graded response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77–107). Routledge/Taylor & Francis Group.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
- Sheehan, K. M. (2013). Measuring cohesion: An approach that accounts for differences in the degree of integration challenge presented by different types of sentences. *Educational Measurement: Issues and Practice*, 32, 28–37. <https://doi.org/10.1111/emip.12017>

- Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, 42(4), 509–524.
- Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, 22(6), 786–809. <https://doi.org/10.1177/0959354312454353>
- Sijtsma, K., & van der Ark, L. A. (2020). *Measurement models for psychological attributes*. CRC Press.
- Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton–Century.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman & Hall/CRC.
- Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Springer.
- Slaney, K. L., & Racine, T. P. (2013). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology*, 31(1), 4–12. <https://doi.org/10.1016/j.newideapsych.2011.02.003>
- Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161–169.
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
- Stein, Z. (2016). *Social justice and educational measurement*. Routledge.
- Stenner, A. J., Fisher, W. P., Jr., Stone, M., & Burdick, D. (2013). Causal Rasch models. *Frontiers in Psychology*, 4, 536.
- Stenner, A. J., Smith, M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20, 305–316.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.
- Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). Wiley.
- Terman, L. M. (1911). The Binet–Simon scale for measuring intelligence: Impressions gained by its application on four hundred non-selected children. *Psychological Clinic*, 5(7), 199–206.
- Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet–Simon intelligence scale*. Houghton Mifflin Co.
- Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, 47(2), 175–186.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4), 567–577.

- Thomson, W. (1851). On a mechanical theory of thermo-electric currents. *Proceedings of the Royal Society of Edinburgh*, 3(42), 91–98.
- Thorndike, E. L. (1904). *Theory of mental and social measurements*. The Science Press. <https://doi.org/10.1037/13283-000>
- Thorndike, E. L. (1905). *The elements of psychology*. A. G. Seiler.
- Thorndike, E. L. (1912). The measurement of educational products. *The School Review*, 20(5), 289–299.
- Thorndike, E. L. (1918). The nature, purposes and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The seventeenth yearbook of the National Society for Study of Education: Part II. The measurement of educational products* (pp. 16–24). Public School Publishing Co.
- Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). American Council on Education.
- Thurstone, L. L. (1925). A method of scaling psychological and educational test. *Journal of Educational Psychology*, 16, 433–451.
- Thurstone, L. L. (1927a). Psychophysical analysis. *American Journal of Psychology*, 38, 368–389.
- Thurstone, L. L. (1927b). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1927c). A mental unit of measurement. *Psychological Review*, 34, 415–423.
- Thurstone, L. L. (1931). *The reliability and validity of tests*. University of Chicago.
- Torres Irribarra, D. (2021). *A pragmatic perspective of measurement*. Springer brief in theoretical advances in psychology. <https://doi.org/10.1007/978-3-030-74025-2>
- Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, 19(5), 579–599.
- Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11, 1–13.
- van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory: Vol. 1. Models*. Chapman & Hall/CRC. <https://doi.org/10.1201/9781315374512>
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.
- von Davier, A. A., Mislevy, R. J., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment*. Springer.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158–177. <https://doi.org/10.1037/h0074428>
- Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). New York, NY: Routledge.
- Wind, S. A., & Hua, C. (2022). *Rasch measurement theory analysis in R*. Chapman & Hall/CRC.

- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. ETS.
- Woodworth, R. S. (1952). Edward Lee Thorndike: A biography. *Biographical Memoirs of the National Academy of Sciences*, 27, 395–428.
- Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 45(4), 51–71.
- Wright, B. D., & Mead, R. J. (1978). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23A). University of Chicago, Department of Education, Statistical Laboratory.
- Yerkes, R. M. (Ed.) (1921). *Memoirs of the National Academy of Sciences, Vol. xv: Psychological examining in the United States Army*. Washington Government Printing Office. <https://doi.org/10.1037/10619-000>
- Zwick, R. (1992). Statistical and psychometric issues in the measurement of educational achievement trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics*, 17(2), 205–218. <https://doi.org/10.2307/1165170>

NOTES

1. In this chapter, we occasionally use the term *construct* instead of *property* when referring to educational and psychosocial measurement, simply for consistency with the vocabulary in these fields. In the present context we treat these terms as synonyms, though with the awareness that they may not always be used synonymously in general (see, e.g., Slaney & Racine, 2013).
2. Some of the territory we cover in this chapter touches on ideas and terminology from the philosophy of science. For a reader looking to gain a greater entry point into these ideas and terminology, we recommend Godfrey-Smith (2003), Maul et al. (2016), Rosenberg and McIntyre (2020), Torres Iribarra (2021), and related entries (e.g., behaviorism, operationalism, realism, pragmatism) that can be found online at the *Stanford Encyclopedia of Philosophy* (<https://plato.stanford.edu/>).
3. A revealing example of the shift from discussion of tests as producing educational measurements to tests as instruments of educational measurement can be found in the name of the organization that is sponsoring this volume, the NCME. NCME was originally founded in 1938 by a group of college professors who were teaching courses on the design and analysis of mental tests. As such, its original name was the National Association of Teachers of Educational Measurements. In 1942, the name was changed to the National Council on Measurements Used in Education. Finally, in 1962, the name was changed to its present form with an emphasis on measurement (i.e., the process), as opposed to measurements (i.e., the products).
4. Here, we are using the term pragmatic in a way that can be viewed as more or less the same as one would use the term instrumentalist.
5. For good introductions, see chapter 2 of Michell (1990) and the first chapter of Krantz et al. (1971). For a retrospective account of representational measurement theory by two of its originators, see Luce and Suppes (2002).

6. Although we have gone through the process of naming each of the two- and three-way interactions depicted in the Venn diagram and identifying the measurement activity that each represents, we do not include it here to keep the discussion at a more general level.
7. <https://www.smarterbalanced.org/assessments/>, retrieved June 25, 2020.
8. See, for example, NAEP's Technology and Engineering Literacy assessment, <https://nces.ed.gov/nationsreportcard/tel/> and <https://nces.ed.gov/nationsreportcard/dba/>.
9. Note that if the "amount" of reading comprehension possessed by a student is only operationally defined by some admissible population of items, the concept of transduction we borrow from thermometry becomes circular since there is no distinction to be made between variability of an attribute and variability of a test score.
10. A red herring argument against the possibility of an interval scale in an IRT context is that the same monotonic transformations of person and item parameters in the 3PLM will leave response probabilities unchanged (Lord, 1980). This is true, but it is irrelevant to the argument that it is the trade-off relationship among persons and items that allows for an interval interpretation.
11. <https://www.nationsreportcard.gov/itemmaps/>

Index

Note: Page references followed by a “t” indicate table; “f”; indicate figure.

- AA-AAAS. *See* alternate assessment aligned to alternate academic achievement standards
- AA-MAS. *See* alternate assessments based on modified achievement standards
- Abdalla, W., 357
- Abdel-Fattah, A. A., 788, 789
- Abdulkadiroglu, A., 1135
- Abedi, J., 119, 121, 135, 145, 237, 974, 1164–65
- Abelson, A. R., 45
- ability estimates, 542, 737, 745–46, 749, 755; adaptive testing and, 82, 353; IRT and, 747–48, 780, 798
- ABMS. *See* American Board of Medical Specialties
- Abrams, L. M., 1035, 1036
- absolute error variance, 302–3, 327, 342–43, 820n1
- Abt, C. C., 520
- academic performance, 119, 1054, 1121, 1165, 1343; ACT relation to, 1225, 1226; GMAT relation to, 1227–28, 1231–32; GRE relation to, 1228–29, 1232; MCAT relation to, 1229–30; test scores and, 426, 1224–25
- accessibility, 408, 983, 992, 1136; in assessment design, 19, 444, 500, 991; through CLR, 963, 966, 968, 978–79, 988–90; computer-based assessment and, 495, 497; for diverse groups, 16, 442; for ELs, 967–68, 974–75; fairness and, 960, 962, 964–67, 999; in ILSAs, 1449; PAD and, 473, 494–95, 982; remotely administered exams relation to, 401; sociocultural context and, 988; state summative assessments and, 1048; for SWDs, 387, 444, 967–68, 971, 975; to TBA, 158; TE and, 984–85; test accommodations for, 609–10, 966, 972, 991; UDA and, 973, 989; UDL for, 394, 961, 973
- Accessibility for Technology-Enhanced Assessments Project, 985
- accountability, 6, 20, 86, 252, 489, 1059; admissions testing and, 231, 1211; culturally responsive assessment relation to, 1168–69; for educational opportunities, 227–28; EOC for, 1139–40; in ESEA, 142; ESSA and, 1118, 1127–30, 1211, 1340; grade promotion exam for, 1140–41; Indigenous populations relation to, 140; K–12 assessment and, 194, 1243; in K–12 education, 21, 483; large-scale testing and, 85; for learning outcomes, 136; in NAEP, 21, 63, 93, 1147–48, 1340; NCLB and, 1049–50, 1118, 1123–27, 1146, 1151, 1156, 1340; PAD for, 445; parents relation to, 1133–35; public, 1130–33, 1138; SLO and, 1190–91, 1240; student achievement relation to, 1143–46; summative assessments for, 442; Title I, 977, 1050, 1119; Tyler rationale and, 1118; unintended consequences of, 1147–48, 1151; validity and, 1188n1; in VAMs, 140. *See also* test-based accountability
- accountability classifications, 1126–27
- accountability data, for state summative assessments, 1136–37
- accountability decisions, 127, 140, 1115, 1116, 1122, 1124
- accountability policies, 1116–17, 1118, 1125–27
- accountability testing, 231, 586–87, 1121; COVID-19 pandemic relation to, 384, 1114; interim assessments for, 1032, 1033, 1037; in K–12 education, 1114, 1115, 1116, 1116f; for special populations, 1163–66; state summative assessments as, 1047; teacher working conditions and, 1147
- accreditation, 833, 879, 1271, 1274, 1320–21; SLO and, 1190, 1237, 1239; training and, 1278
- Accreditation Council for Graduate Medical Education, 1320–21
- ACCUPLACER, 1234, 1236–37
- Achieve, 231, 1047, 1139, 1157–58
- achievement gaps, 1117, 1126, 1144–46; SES relation to, 231, 977–78; summative assessments and, 1051
- achievement-level descriptors, 1169, 1546–47
- achievement levels, 460, 722, 1048, 1272, 1554; cut scores and, 927, 1556, 1557; validity and, 210, 1132
- achievement tests, 86, 384, 444, 765, 978; fairness of, 397–98; grading of, 1043; standardized tests for, 386; test scores for, 200; validity and, 201, 1193
- Ackerman, T. A., 1515
- Ackerson, L., 45
- acquiescence response style (ARS), 1365
- ACT, 88, 1204–6, 1264n10, 1356; academic performance relation to, 1225, 1226; accountability and, 1211; for accountability testing, 231; alignment and, 1159; battery scaling and, 792–93; concordance and, 788–89; COVID-19 pandemic and, 1191, 1212, 1213; decision-making and, 133–34; equating and, 766; flagging policies for, 405–6; grading and, 1043; HSEEs and, 1139, 1140; HSGPA relation to, 1197–98, 1226; linking invariance and, 419, 420; norms and, 758; PBA and, 600; as placement test, 1234; SAT compared to, 112n30; scale scores for, 750, 752, 754; SEM and, 757; social justice relation to, 1192; TBA and, 586, 607; test accommodations for, 968; test scores on, 1211; UC and, 406–7, 1193, 1194; validity of, 1231; vertical scaling and, 789
- ACT and SAT Content, Testing Time, and Item Counts, 1207t
- ACT Aspire, 1158, 1159
- ACT Composite score, 1206, 1208
- ACT Enhancements, 1206
- ACT Holistic Framework for Education and Work Success, 494
- ACT-R models, 628–29
- ADA. *See* Americans With Disabilities Act
- Adams, C. J., 1211
- adaptive learning, 623, 632; Bayesian models for, 628–29; validity and, 630
- adaptive testing, 1482; ability estimates and, 83, 353; digital assessments and, 1450–51, 1456; GRE and, 610, 1218; ILSAs and, 1447–48; IRT, 765, 797–98; linking and, 772, 798–99; proficiency estimates in, 626; reliability coefficients for, 352, 366–67; UIRT and, 799–800. *See also* computerized adaptive testing (CAT)
- additive conjoint measurement, 1532–33, 1553, 1561
- adequate notice, 150, 397, 1150

- adequate yearly progress (AYP), 143, 978, 1035, 1123, 1124
- Adewuya, A. O., 127
- Adler, J., 135
- administration conditions, 797, 1300–1301; fairness and, 409; GT and, 356; in large-scale testing, 124; reliability and, 364; scaling and, 737; standardization and, 242, 249, 607; test accommodations and, 516–18; validity and, 21, 1114
- administrative records, 1357, 1390, 1391
- admissions testing: accountability and, 231, 1211; assessment design of, 1222; coaching for, 396–97; consequences of, 231–32; content standards relation to, 231; COVID-19 pandemic and, 15, 22, 400, 1191, 1212, 1213; diversity and, 1243; equitability of, 14; fairness in, 21–22, 387, 404–5, 425, 426, 1242; GPA and, 417; for graduate school admissions, 1213; language testing programs in, 386; percentile-rank system in, 472; prediction invariance and, 419; reliability of, 21–22, 1223, 1223*t*, 1224; scaling in, 737; SES and, 393, 407–8; social justice and, 22, 1192; Sorbonne entrance examination, 385; standardization of, 426; standardized tests for, 134; as state testing, 231–32; test accommodations for, 971–72; test-optional policies and, 423; test scores for, 1190, 1195; TOEFL for, 234; for UC, 406–8; for undergraduate admissions, 1210–11; validation of, 1199, 1201; validity and, 21–22, 1223, 1224–25, 1225*t*. *See also specific tests*
- Advanced Placement (AP), 442–43, 455; for African American studies, 6–7; construct definition and, 446–47; gender DIF and, 987; Music Theory Exam, 528; NRC relation to, 446; PAD and, 500; score weights for, 744; SEs and, 781; test accommodations for, 401
- AERA. *See* American Educational Research Association
- AES. *See* automated essay scoring
- affect heuristic, 846–47, 851, 875
- affirmative action, 1200
- Agger, C., 877
- Agrawal, N., 405–6
- Ahn, T., 1144
- AI. *See* artificial intelligence
- Ai, F., 628
- Aicheson, J., 1376
- AICPA. *See* Association of International Certified Public Accountants
- AIG. *See* automated item generation
- Aikens, H. A., 43
- Ainley, J., 1510n55
- Ainley, M., 1510n55
- Ainsworth, L., 1038, 1039
- Airasian, P. W., 1138
- Air Force School of Aviation Medicine, U.S., 60, 77, 1125
- Aitkin, M., 664, 686, 687–88, 1534
- Akaike, H., 677, 678
- Akaike information criterion, 678
- Alan, S., 1387
- Alarifi, M., 946
- Alavi, S. M., 986
- Alayed, E. F., 520
- Albano, A. D., 985
- Albano, A. M., 127
- Albert, J. H., 663
- Alderson, J. C., 234
- ALEKS. *See* Assessment and Learning in Knowledge Spaces
- Aleven, V., 629, 630
- Alexander, L., 1131
- Alexander, P., 1542, 1543
- Ali, A. A., 1365
- alignment: cognitive complexity and, 1160; content validity and, 489; in EOC, 395; linear scale, 765, 784–85; post hoc, 493; RPLDs and, 492; score reporting relation to, 908; state testing and, 493, 1157–58, 1159; of test specifications, 389; test-to-standards, 1159; validity and, 1156–57
- Alim, 497
- Allal, L., 1022
- Allalouf, A., 983
- Allen, A., 1053, 1055–56, 1059
- Allen, J., 1206, 1208, 1211, 1225, 1226, 1231
- Allen, N. L., 759, 916, 920, 927
- Allensworth, E., 1137
- Alliance for Excellent Education, 1129
- Almlund, M., 1342–43, 1353
- Almond, R., 617, 630
- Almy, S., 1141
- Alonzo, A. C., 1063, 1068, 1069, 1075
- Alreck, P. L., 1358
- alternate assessment aligned to alternate academic achievement standards (AA-AAAS), 970–71, 973
- alternate assessments: performance standards for, 970; for SWDs, 511n2, 832, 972
- alternate assessments based on modified achievement standards (AA-MAS), 970
- alternate English language proficiency assessment (Alt ELPA), 718, 719*t*, 722
- alternative online administrations, 18
- Alvero, A. J., 131, 1199
- Alves, 1305
- Alzahrani, S. M., 520, 521
- Amalnerkar, E., 1300
- ambulatory assessments, 1389
- American Association for the Advancement of Science, 1121
- American Association for the Study of the Feeble-Minded, 52
- American Association of Medical Colleges, 1382
- American Bar Association, 1215, 1219
- American Board of Anesthesiology, 1322
- American Board of Internal Medicine, 1274, 1294, 1295–96
- American Board of Medical Specialties (ABMS), 1276, 1295, 1321–22
- American College Testing, 88, 112nn30, 1204
- American Community Survey, 972
- American Council on Education, Cooperative Test Service, 58
- American Educational Research Association (AERA), 79, 121, 123–24, 151, 192, 194, 197, 198, 203, 204, 216, 222–23, 227, 330, 337, 361, 362, 386, 387, 388, 389–90, 391–92, 394, 395, 400, 405, 406, 419, 426, 437n4, 443, 514, 550, 556, 583, 597, 605–6, 737, 804, 822, 840, 853, 905, 906, 907, 908, 910, 921, 925–26, 960, 961, 962–63, 968, 969, 973, 988–89, 990, 991, 992, 994, 997, 998, 1020, 1116, 1117, 1119, 1150, 1154, 1188n1, 1222, 1230, 1233, 1274, 1279, 1282, 1287, 1288, 1298, 1310, 1312, 1316, 1317–18, 1319, 1363, 1388, 1437, 1549, 1550
- American Federation of Teachers, 147
- American Institutes for Research, 1128
- American Institutes of Certified Public Accountants, 586
- American Legion, 5
- American Library Association, 8
- American Psychological Association (APA), 52–53, 199, 216, 389–90, 867
- American Recovery and Reinvestment Act (2009), 587
- American Society of Association Executives, 1322
- American Statistical Association, 1154
- Americans With Disabilities Act (ADA), 517, 1280; civil rights in, 142; test accommodations and, 135, 139, 142, 971, 1283–86
- Ames, C., 1022
- AMO. *See* annual measurable objective
- Amrein, A. L., 148
- An, J., 349
- analytical judgment, 849–50
- Anastasi, A., 353, 354, 449
- anchoring and adjustment bias, 847
- anchoring methods, 23, 1371
- anchoring vignettes (AVs), 1371–74, 1371
- anchor test nonequivalent groups design, 605, 766
- Andersen, E. B., 77
- Anderson, C. A., 846
- Anderson, J. I., 62
- Anderson, J. R., 623, 628
- Anderson, L. W., 459, 1043, 1044
- Anderson, N., 133, 401, 1211
- Andersson, B., 313, 319, 360
- Andes, 627
- Andrade, H. L., 485, 1020, 1022, 1029–30, 1031
- Andrew, B. J., 1316

- Andrews, J. J., 260
- Andrich, David, 81, 525, 740, 747, 1360, 1378, 1532, 1533
- Andrienko, G., 944
- Angoff, W. H., 60, 82, 91, 112n21, 292, 404, 736, 742, 749, 758, 760, 762, 763, 765, 766, 770, 773, 776, 777, 778, 781, 782, 786, 788, 801, 802, 838, 841, 1311, 1547
- Angoff-Feldt coefficient, 292–93
- Angoff method, 841, 1314; borderline test takers' relation to, 866, 867, 869; cognitive-judgmental task in, 873; for credentialing, 1311–12; cut score in, 870; extended version, 826, 850, 866; modified version, 829, 835, 838, 865–67, 870–71, 1311; OIB and, 844; panelists, relation to, 865–67, 872–73; yes/no version, 826, 835
- Angoff probability judgments, 848–49, 850
- L'année psychologique* (journal), 48–49
- Annetta, L. A., 520
- annual measurable objective (AMO), 1123, 1125
- ANOVA. *See* application of analysis of variance
- Anrig, Greg, 403
- Ansari, A., 1128
- Ansley, T., 931
- Anthony, L. C., 1229
- anthropometric measurement, 51
- antiracist framework, 11, 996
- antiteasing sentiments, 23, 53, 57, 1268, 1270, 1319
- AP. *See* Advanced Placement
- APA. *See* American Psychological Association
- AP Music Theory Exam, 528
- application of analysis of variance (ANOVA), 297–98, 317; error variance and, 313, 319; in G studies, 300; intraclass correlation coefficient and, 311–12; multivariate GT and, 305
- Applications of Item Response Theory to Practical Testing Problems* (Lord), 76
- Applied Measurement in Education* (journal), 1309
- Architect Registration Examination (ARE), 599, 1287–88, 1294
- argument-based approach, 192, 208; to validation framework, 215, 482–83; to validity, 451, 909–10
- Argyle, E. M., 946
- Aristotle, 12, 92
- Armed Services Vocational Aptitude Battery (ASVAB), 61, 585
- Armor, D. J., 1134–35
- Armstrong-Carter, E., 127
- Army Air Forces, 59–61
- Army Alpha, 53–55, 58, 73, 898, 1522; equating and, 764; eugenics movement and, 53, 55–56; MC and, 87, 92; National Intelligence Tests and, 386
- Arneson, A., 1069
- Aronson, J., 154
- ARS. *See* acquiescence response style
- Arthur, W., 1382–83, 1386
- artificial intelligence (AI), 131, 622; for assessment design, 493–94; in assessment development, 137; automated scoring by, 130, 358, 618–19, 635–36; CAI with, 623; in games and simulations, 624, 637; LLMs, 555, 619–20, 637; as raters, 358–59, 542; reliability of, 551; in remotely administered exams, 400; rubrics relation to, 534; test scores and, 532, 533; translation procedures relation to, 1443; validity and, 550
- Artiles, A. J., 147
- Arundel, K., 1206
- Ashton, M. C., 1352, 1356
- Aspen Institute, 1340
- Asplund, J., 1376
- ASR. *See* automated speech assessment
- Assessment and Learning in Knowledge Spaces (ALEKS), 623, 636
- assessment as social phenomenon, 125–27, 129
- assessment capacity, 138
- assessment constraints, 455–56; PAD and, 459–60; RPLDs and, 463
- assessment design, 16–17, 443, 597; accessibility in, 19, 444, 500, 991; accountability for, 489; for admissions testing, 1222; AI for, 493–94; for Alt ELPA, 722; bias in, 444, 494, 497; CBAL and, 219; for CCSS, 455; of classroom formative assessments, 1023; CLR in, 235, 258, 494–95, 990, 991; cognition relation to, 456–57, 458, 460; of common formative assessments, 1040–41; conditional inference in, 165; consequences of, 228; construct-irrelevant variance and, 454–55; of culturally responsive assessment, 243–44; culture relation to, 235, 242–43; data collection for, 775–76; of digital assessments, 1440, 1451–52; diversity relation to, 149, 473–74; domain analysis for, 454; evidentiary argument in, 485–86; fairness relation to, 236, 961, 998–99; games and simulations relation to, 257; of ILSAs, 1438; Improving America's Schools Act relation to, 442; for interim assessment, 1033; item analysis relation to, 705; for K–12 education, 493; language in, 237; learning progressions in, 1063–64, 1068, 1071; OTL and, 478, 479; performance inferences in, 212; PLDs in, 877, 878; procedural validity evidence for, 482, 486, 487; psychometrics in, 260, 479–80, 481; Remaining Steps in, 465f; response process and, 254–55; RPLDs and, 471, 476–77, 483, 489; score reporting relation to, 946; sensitivity guidelines for, 395–96; sociocultural context and, 120, 125, 166–68, 444; with sociocultural learning theories, 982–83; standardization of, 249; of state summative assessments, 1047; TBA and, 583, 590, 635; test-taker characteristics relation to, 241; UX and, 498; validation and, 480, 723; validity and, 192, 194, 205, 208, 613. *See also* performance-level descriptors; principled assessment design
- assessment development, 129, 130, 597; of admissions testing, 1222; AIG and, 1292; AI in, 137; assessment literacy in, 1080; of background questionnaire, 1436–37; bias in, 1269; for certification, 1277; CLR and, 997, 999; for credentialing, 1287–89; cultural validity in, 992–93; data collection for, 1290–91; DIF and, 414, 988; documentation and, 485–86; educational disparities relation to, 980; ELs relation to, 976; fairness and, 598, 961, 982, 983; feedback in, 1338n4; of ILSAs, 1440–41; IUA and, 907; for K–12 assessment, 994; PAD and, 452, 472–73, 983; reliability and, 767; score interpretation in, 929–30; score reporting relation to, 898, 946; sociocultural context in, 988; stakeholders relation to, 1268; UDA in, 973; UDL in, 983, 1166; validation framework in, 199
- assessment engineering, 445–46, 909–10
- assessment for teaching and learning, 1018, 1078; assessment literacy and, 1080; formative assessments relation to, 1020–21; state summative assessments as, 1047, 1051
- assessment interpretation: fairness in, 136; of large-scale testing, 194–95; social inequities in, 138; sociocultural context in, 118–20, 121, 130; validity of, 131, 192–93, 197, 204, 391
- assessment literacy, 138, 1019, 1027, 1076–77; assessment for teaching and learning and, 1080; CBM and, 1059; common formative assessments and, 1041; grading and, 1045; performance assessment and, 1040
- Assessment of Higher Education Learning Outcomes, 1238
- assessment of learning spans or trajectories, 1052
- assessment performance, 1034; construct-irrelevant variance relation to, 391–92; culture relation to, 236–37; fairness and, 134–35; sociocultural context relation to, 123, 139, 156–57, 165
- assessment quality, 16, 442, 445
- Assessment Reform Group, 1022
- assessment tasks, cultural relevance of, 121
- The Assessment Triangle, 152f, 160f, 454. *See also* cognition model; interpretation model; observation model
- assessment use argument (AUA), 232, 232f
- Assessment Work Group, 1350
- ASSISTments, 627

- Association of American Medical Colleges, 1222, 1229, 1230
Association of American Medical Colleges (AAMC) v. Carey (1990), 404–5
 Association of Black Psychologists, 977
 Association of International Certified Public Accountants (AICPA), 1320
 Association of Test Publishers, 448, 550, 1293, 1294
 ASVAB. *See* Armed Services Vocational Aptitude Battery
 Atchison, D., 1144
 Atitya, B., 1033
 Atkin, J. M., 1484
 Atkinson, D., 226
 Atkinson, R., 385
 Atkinson, Richard C., 1203
 Attali, Y., 354, 358, 359, 524, 534, 535–36, 620
 Attewell, P. A., 1233
 Atwell, M. N., 1340
 Au, W., 977, 1050
 AUA. *See* assessment use argument
 audience-based research, 902, 903, 903f
 audience identification, 930
 audio prompts, 528
 Auerbach, M. A., 1311
 augmented subscores, 920
 August, D., 973, 978
 Australia, 1486
 authoritative feedback, 1027
 autoencoder, 549
 automated essay scoring (AES), 15, 532, 533–34, 535f, 561; ASR compared to, 539; foreign-language scoring engines and, 559, 559f; Hewlett trials and, 556–57; LSA for, 545; Turing test and, 550; validity and, 255–57
 automated item generation (AIG), 1292, 1305–6
 automated scoring: by AI, 130, 358, 618–19, 635–36; bias in, 137, 391, 399; of CR, 584; decision/use inferences for, 257; feature extraction for, 542, 546, 547; human raters compared to, 358–59, 538, 558, 562, 620; ILSAs with, 1482; Kaggle platform and, 557; for kinesthetic assessment, 552; for licensure, 586; linking and, 795; neural networks for, 749; PEG for, 554, 554f, 555; response process and, 653n2; topic models and, 550; training and, 521, 536; validity of, 15, 255–56, 620–22
 automated speech assessment (ASR), 539, 539f, 540–42
 automated video interviews, 552–53
 automation, 1242
 Autor, D. H., 1344
 autoregressive time series modelling, 350
 Autoscore, 547
 auxiliary scales, 738, 793; norms and, 757–58; primary scales with, 759
 Avery, C., 1211
 Aviation Cadet Qualifying Examination, 59
 Aviation Psychology Program, 59
 AVs. *See* anchoring vignettes
 Avvisati, F., 1509n48
 AYP. *See* adequate yearly progress
 Azuma, Y., 349
 Babo, G., 1034
 Baca, G., 148
 Bach, B., 911, 912, 913
 Bacher-Hicks, A., 1143, 1154
 Bachman, L. F., 219, 220, 228, 232, 233, 275n4
 Backes, B., 606
 background characteristics, 658, 1427, 1452
 background questionnaire, 24, 1388–89, 1439, 1444, 1472; assessment development of, 1436–37; ILSAs and, 1441–42, 1490; matrix sampling design and, 1446; measurement invariance and, 1456–57; OECD and, 1351; PIAAC and, 1431–32, 1434; population model and, 1452; relevance and, 1438, 1481; scaling and, 1451, 1455, 1459, 1478–79; TBA and, 1490
 Bäckström, M., 1378
 Bacon, D. R., 349
 Bagley, William C., 42
 bag-of-words approach, 555, 616
 Bailey, A. L., 149, 1029, 1067
 Bailey, D., 1341
 Bailey, J. M., 1042
 Bailey, T., 1233–34
 Baird, J.-A., 1070, 1469
 Bakeman, R., 1069
 Baker, B., 137
 Baker, C. L., 617
 Baker, E. L., 1148, 1149, 1154
 Baker, F. B., 1529–30, 1534
 Baker, O., 1148
 Baker, R. S. J. d., 616, 626, 928
 Baker-Bell, A., 996
 Baki, O., 127
 balanced incomplete block (BIB), 1433, 1446, 1489
 Baldwin, P., 803, 868–69
 Ball, D. L., 1040
 Balloo, K., 1031
 Balthazard, C., 1266
 Bancroft, K., 1035
 Bandeira de Mello, V., 802, 1132
 Bandura, A., 1355
 Banker, H. J., 1042–43
 Banks, J., 529
 Banks, K., 120, 986
 Baradwaj, B. K., 616
 Barber, Cosmetology/ Barber, Esthetics, Electricity and Nail Technology Licensing Board, 1272
 Bargmann, R. E., 712
 Barnard, G. A., 77
 Barnes, L. L., 1074
 Barnes, M., 234
 Barnett, E., 1234
 Barnett, S., 1341–42
 Barnhart, R., 80
 Baron, R. J., 838, 1319
 Barone, J. L., 1432
 Barrick, M. R., 1342
 Barrington, 1343–44
 Barron, S. I., 875, 1160
 Barrow, L., 1341
 BARS. *See* behaviorally anchored rating scales
 Bartholomew, D. J., 660, 683, 694, 696
 Bartlett, B., 1156, 1157
 Barton, D., 1344
 Bartram, D., 1355, 1376, 1380–81
 Bashkov, B. M., 1291
 basket standard-setting method, 871, 894n11
 Bass, A. R., 416, 421
 Bastedo, M. N., 423
 battery scaling, 756–57, 792–93
 Bauer, D. J., 686, 701, 721–22
 Bauer, G. R., 986
 Bauer-Wolf, J., 1264n11
 Baxter, G. P., 1074
 Baxter, J. C., 1366, 1368
 Bay, L., 1237
 Bayer, P., 1135
 Bayesian conceptualization, 1535
 Bayesian estimation, 34, 284, 317, 663, 686
 Bayesian framework, 310; CTT in, 64; error variance in, 329; IRT and, 309; latent variable models and, 670; MAP and, 317; Markov chain and, 347; population model and, 1452; true scale score in, 323
 Bayesian information criterion, 678
 Bayesian knowledge tracing (BKT), 624, 627, 628
 Bayesian models, for adaptive learning, 628–29
 Bayesian networks, 598–99, 629, 629f, 630, 1320; probabilistic modeling in, 623; response process and, 635; theory-driven modeling and, 617
 Bayesian posterior predictive model checking, 697–98
 Bayes model estimate, 352
 Bayes nets, 157
 Bazerman, C., 553–54
 Bazerman, M., 841
 Beach, K., 158
 Bearak, J. M., 126, 1147
 BEAR Assessment System, 1064–65, 1066, 1069, 1070
 Beard, J. J., 88
 Beaton, A. E., 916, 920, 927, 1432, 1548, 1561
 Beatty, A. S., 1043, 1210, 1242–43
 Bechard, 970
 Becker, D., 1380
 Becker, G. S., 1341
 Becker, K. A., 119, 739, 855
 Bedard, K., 1474
 Beddow, P. A., 964–65, 966, 982, 994

- Beguín, A., 1163
 behavioral learning theories, 1542, 1543–44
 behaviorally anchored rating scales (BARS), 1371, 1374–76, 1458
 behavioral measures, 1389
 behavioral observations, 359–60
 behavioral summary scales (BSS), 1371
 Behrens, J. T., 500, 911–12
 Beier, M. E., 1384
 Beigman-Klebanov, B., 535
 Beiser, D., 352
 Bejar, I. I., 130, 131, 256–57, 460, 476, 483–84, 550, 586, 599, 619, 637, 859, 877
 Belasco, A. S., 1196
 Bell, 1018
 Bell, Terrell, 1121
The Bell Curve (Murray), 57
 Bellman, R., 617
 Bellwether Education Partners, 1129
 Benavot, A., 1425, 1484
 benchmark assessments, 1032, 1037
 Benchmarked Standard Setting approaches, 877
 benchmarking, 1471, 1486, 1488
 Bender, M. C., 27
 Benet-Martínez, V., 1342–43
 Benitez, I., 614
 Bennett, C. T., 1230
 Bennett, R. E., 17, 87, 116, 125, 165, 219, 254, 256, 484, 550, 582, 583, 586, 587–88, 590, 591, 597, 608, 614, 617, 618, 620–21, 631, 632, 633, 637, 660, 25, 984, 995, 1018, 1019, 1049, 1071, 1072, 1073, 1131, 1168–69, 1195, 1196, 1293, 1388
 Bentler, P. M., 678
 Benzeghiba, M., 137
 Beretvas, S. N., 681
 Berg, J., 1356
 Bergman, M. E., 1384
 Bergner, Y., 660, 946, 1356, 1459
 Berk, R. A., 338
 Berkeley Evaluation and Assessment Research Assessment System, 1551
 Berkson, Joseph, 77
 Berliner, D. C., 148
 Berman, A. I., 119, 194, 602
 Bernardin, H. J., 522, 524
 Bernoulli distribution, 671
 Bernstein, I. H., 1360
 Berry, C. M., 1226
 BERT. *See* Bidirectional Encoder Representations from Transformers
 Bertling, J. P., 1358, 1373, 1479
 Bertrand, M., 1136
 best-choice instructions, 1383
 beta-binomial model, 340–41, 343
 Betebenner, D. W., 793, 1152, 1153–54
 Betts, J. R., 393
 Bezirhan, U., 1457
 BF-MIRT. *See* bifactor MIRT
 Bhat, S., 541
 Biancarosa, G., 561
 bias, 16, 91; ACT relation to, 1231; in assessment design, 444, 494, 497; in assessment development, 1269; in automated scoring, 137, 391, 399; in coaching, 1162; comparability relation to, 605; confirmation, 209, 216, 846; in CSEM, 325, 337; culture relation to, 138–39; data smoothing relation to, 773, 774; in decision-making, 148, 841–42, 846; DIF and, 985; discrimination and, 141; diversity and, 235; in employment tests, 403; equal performance principle and, 392; expected error variance and, 282–83; fairness and, 163, 423, 436n2, 961, 1308; homogeneity assumption and, 123; in human judgment, 823, 834; of human proctors, 402; in intelligence testing, 50, 56; IRT and, 329, 672, 1365; item, 404, 411; IUA relation to, 261; in large-scale testing, 121, 133–34, 993–94; latent proficiency variance and, 314; linguistic heterogeneity, 119; MLE and, 310, 315; nonresponse, 1445; in odd-even split, 293; Personal Potential Index and, 1370; in PIAAC, 1456; in PISA, 1472; PLDs and, 847; population homogeneity and, 122; precision and, 364–65; predictive, 388, 417; in probability judgments, 838, 846; proficiency estimates and, 353; racial, 259; of raters, 246, 523, 555, 796; regression invariance and, 416–17; reliability coefficients and, 290–91, 315, 360; research agenda for, 875; RSM and, 1391; scale-aligning methods and, 784; scale anchoring and, 1371; score inferences relation to, 213; in selection decisions, 421; self-selection, 553; social-desirability, 1362; sociocultural context and, 158; in SSAT-II, 397; standardization and, 609; in standard setting, 846, 1319; standard-setting design and, 876; in test interpretation, 997; training for, 860, 861, 862*t*; in translation procedures, 1479; validity relation to, 132–33, 390, 1308; VAMs and, 351; variance and, 312–13
 BIB. *See* balanced incomplete block
 Biber, D., 561
 BICAL, 1534
 Bich, W., 1539
 Bickhard, M. H., 1555
 BICSE. *See* Board on International and Comparative Studies in Education
 Bidirectional Encoder Representations from Transformers (BERT), 536, 538, 555
 The Bidirectional Relationship Between Sociocultural Context and Assessment, 117*f*
 Bieber, T., 140
 Biemer, P., 1445
 Bieri, J., 842
 bifactor MIRT (BF-MIRT), 320–21, 345
 bifactor models, 698, 707, 709*t*, 711; DCM compared to, 710; multilevel, 708, 709, 718; testlet model and, 666–67; UIRT *versus*, 708*t*
 big data, 137, 944, 946, 1390–91, 1566
 Big Five personality model, 1342–43, 1346, 1348; CASEL compared to, 1356–57; GPA relation to, 1368–69; personality tests and, 415, 1356; psychometrics of, 1352; self-report ratings and, 1358
 Biggs, J. B., 459
 bigram topic model, 549
 bilingual assessments, 974–75
 Billingham, C. M., 1135
 Binet, Alfred, 34, 35–36, 37, 47–50, 93, 110n8, 132, 385, 739, 741, 759, 1520–21, 1522
 Binet–Simon Intelligence Test, 385, 739, 741, 759
 Binet–Simon scale, 49–50, 52, 1523
 Bingham, R. O., 1213, 1222
 Bingham, Walter, 34
 Binkley, C., 1198
 Binkley, M., 1434, 1482
 binomial error model, 326–27, 333
 biodata, 1389–90, 1391
Biometrika (journal), 41
 Birnbaum, S., 7
 Birnbaum, Allan, 60, 76–77, 78, 90, 91, 665, 667, 1378, 1532, 1548
 Bishop, Y. M. M., 692, 696
 bivariate normal distribution, true score and, 342
 Bjork, R. A., 1025
 Björklund, F., 1378
 BKT. *See* Bayesian knowledge tracing
 Black, A. W., 533
 Black, P., 250, 1020, 1021, 1030, 1065, 1069, 1075, 1484
 Black, S. E., 1135
 Blackburn, J., 520
 Black Lives Matter, 8–9, 30n3, 1191–92
 Blanc, S., 1034–35, 1036
 Blatter, A., 423
 Blazar, D., 1141
 Blei, D. M., 542–43, 545, 546, 548
 Bleiberg, J., 1143
 Blinder, A., 1148
 blind validity checks, for raters, 523–24
 Bloom, B. S., 1019–20, 1056, 1119, 1120
 Bloom's taxonomy, 459
 Blyth, D. A., 1350, 1356, 1357
 Board on International and Comparative Studies in Education (BICSE), 1432, 1507n11
 Bock, R. Darrell, 80, 81, 83, 316, 329, 664, 686, 687–88, 712, 1359, 1360, 1534
 Bock–Aitkin EM, 688–89, 703
 Böckenholt, U., 1362
 Bocklisch, F., 1359
 Bodily, R., 924
 Body of Work method, 827, 843, 845

- Boeck, 662
 Boekaerts, M., 1030
 Bohrnstedt, G., 118, 977, 978
 Bok, H. G. J., 1209, 1321
 Bollen, K. A., 663, 676
 Bolt, D. M., 163, 164, 802
 Bolt, S., 968
 Bolt, S. E., 517
 Bond, T. G., 1533, 1547
 Bonferroni method, 437n7
 Bonifay, W., 656, 698, 709, 711
 Bonilla, S., 1144, 1168
 Bonito, J. A., 832
 Booher-Jennings, J., 228–29, 1137, 1148
 book bans, 8
 Bookmark method, 225, 842, 849, 865, 872;
 borderline test takers relation to, 869;
 cognitive-judgmental task in, 867–68,
 873; OIB in, 894n9; panelists relation to,
 867–68, 871; for principal evaluation, 870;
 Smarter Balanced Assessment Consor-
 tium and, 855
 Boolean variables, 627, 628
 Boost, I., 1198
 Borderline Group method, 837t, 841
 borderline test takers, 838, 861, 893n1; Angoff
 method relation to, 866, 867, 869; cog-
 nitive complexity and, 842–43; panelists
 relation to, 869; standard-setting design
 relation to, 871–72; threshold-signal
 detection judgments and, 839–40
 Borghans, L., 1343
 Boring, E. G., 46
 Borman, W. C., 1371
 Bornstein, M. H., 1365
 Borreca, C. P., 142, 144, 149
 Borsboom, D., 162, 209, 218, 1514, 1528,
 1531, 1549, 1550–51, 1552, 1553, 1555
 Boser, U., 1140
 Boshara, R., 1209
 Boston public schools, 84
 Boud, D., 1026
 boundary objects, 122
 bounded rationality, 841
 Bourdieu, P., 130
 Bourque, M., 866
 Bowen, W. G., 1209
 Bowers, A. J., 1043
 Bowers, J. J., 1312
 Bowles, S., 1341
 Box, C., 1029, 1071
 Box, G. E., 678
 box plot, 658
 Boyd, B., 350
 Boyd-Graber, J., 546
 Bracey, G. W., 1472
 Bradburn, N. M., 1432
 Bradley, R., 1057
 Bradshaw, L., 348, 350–51, 1075
 Bränberg, K., 801
 Branch, G., 7
 Brandmaier, A. M., 350
 Brandon, P., 851
 Bransford, 153
 Bratko, D., 1346
 Braun, H., 23, 134, 586, 587, 599, 619, 659,
 679, 765, 777, 781, 802, 1051, 1115, 1240,
 1425, 1433, 1436, 1467, 1471, 1473, 1484,
 1485, 1489
 Bravo-Senzana, M. V., 1340
 Breakspear, S., 139, 1468
 Breil, S. M., 1389
 Breiman, L., 656, 724
 Breimhorst, Mark, 405
Breimhorst v. Educational Testing Service
 (2000), 405–6
 Breithaupt, K. J., 586
 Brennan, R. L., 70, 90, 111n19, 197, 213, 278,
 279, 285, 288, 291, 292, 294, 295, 297, 298,
 304–5, 306, 307, 324, 328–29, 332, 335,
 338, 340, 343, 344–45, 347, 350, 351, 355,
 360, 603, 605, 722, 737, 740, 743, 745,
 752–53, 754, 757, 759, 760, 765–66, 768,
 769, 770, 771, 772–73, 774, 775, 776, 778,
 779, 780, 781, 783, 785, 789–90, 792–93,
 794, 795, 796, 803, 804, 820n1, 897, 1298,
 1316, 1367, 1528, 1547, 1555, 1558
 Breyer, F. J., 338, 1298
 Bridgeford, N. J., 534
 Bridgeland, J. M., 1340
 Bridgeman, B., 137, 233, 388–89, 398, 556,
 653n4, 1226, 1228
 bridge studies, 608
 Briggs, D. C., 24, 70, 71, 75, 124, 739, 741, 742,
 746, 747, 748, 755, 789, 791, 839, 1060,
 1068, 1075, 1518, 1519, 1520–21, 1522,
 1529, 1540, 1547, 1553, 1560, 1564
 Brigham, C. C., 56, 87, 111n14, 385, 386
 Brightpath, 924
 Brillinger, D. R., 660
 British Association for the Advancement of
 Science, 1521
 British Indian civil service, 35, 37, 83
British Journal of Mathematical and Statistical
Psychology (journal), 1365
British Journal of Psychology (journal), 44
 Broadbuss, A., 1067, 1075
 Broadfoot, A. A., 1383
 Brogden, J. E., 1533
 Bromwich, J. E., 402
 Bronak, S. C., 520
 Brookhart, S. M., 20, 118, 386, 485, 622, 657,
 926, 962, 998, 1022, 1023, 1026, 1029,
 1041, 1043, 1044, 1045, 1115
 Brooks, J. L., 38
 Brossman, B., 1296
 Broussard, M., 749
 Brown, A., 1376–77, 1379, 1380
 Brown, G. T. L., 132, 924, 925, 926, 942, 943,
 946
 Brown, J. S., 444, 623
 Brown, M. L., 1063
 Brown, R. S., 1034
 Brown, T., 605
 Brown, W., 44–45, 291
 Browne, M. W., 676–77, 678
 Browne, W. J., 681
 Bruning, R., 1025
 Brunza, J. J., 363
 Bryan, W. J., 55
 Bryk, A. S., 669, 679, 681, 701, 1049
 BSS. *See* behavioral summary scales
 Buchanan, R., 1045
 Buckendahl, C. W., 22, 853, 864, 872, 1311,
 1313–14
 Buckley, K., 522, 1060
 Buckley, K. H., 1142
 Buell, J. Y., 1069
 Bujang, M. A., 360
 Bulkley, K. E., 1035
 Bulut, O., 943, 944
 Bunch, M. B., 62, 822, 852, 856, 859, 860, 861,
 1310, 1547
 Bunderson, C. V., 1072, 1078, 1079
 Bunji, K., 1380
 Burch, P. E., 1032, 1136
 Burdett, N., 1467, 1468, 1471–72, 1484
 Bureau of Applied Social Research, Columbia
 University, 76
 Burkam, D. T., 393
 Burke, M., 749
 Burkhardt, H., 1031
 Bürkner, P.-C., 1379
 Burns, D., 243–44
 Burnstein, E., 831, 863
 Buros Mental Measurements Yearbook, 1033,
 1346
 Burrus, J., 1344
 Burstein, J., 534, 535–36, 539, 583, 619
 Burstein, N., 135, 358
 Burt, Cyril, 69, 69, 298
 Burt, W. M., 872
 Burton, E., 598
 Burton, N. W., 1232
 Busch, J. C., 869, 874, 1312
 Busche, K., 1230, 1233
 Bush, G. H. W., 1131
 Bush, G. W., 977–78
 Butler, D. L., 1022, 1025
 Buxton, C., 517
 Buyse, T., 1382
 Buzick, H., 240, 969, 970
 Byrne, B., 1456
 CAASPP. *See* California Assessment of
 Student Performance and Progress
 Cadle, A., 1293
 Caesar, J., 974
 Cahill, A., 1366
 CAI. *See* computer-assisted instruction
 Cai, L., 17, 320, 656, 659, 665, 666, 669, 673,
 675, 678, 683, 685, 686, 689–90, 691, 694,
 696, 697, 698, 701, 703, 705, 706, 707, 708,

- 709, 711–12, 722, 745, 1075, 1320, 1365, 1391, 1524, 1535
- Caines, J., 870, 874
- calibration, 1454, 1459; concurrent, 605; faking and, 1362, 1368, 1370–71; IRT and, 162, 605, 658, 722, 745, 791, 1548; item, 1451; Likert scale and, 1378; linking and, 786; qualification exam and, 523
- California Assessment of Student Performance and Progress (CAASPP), 582
- California Critical Thinking Assessment, 1239
- California Office to Reform Education (CORE), 1340, 1354–55, 1359, 1362–63, 1367
- Camara, W. J., 5, 6, 21, 149, 385, 393, 795, 962, 1043, 1115, 1191, 1193, 1194, 1195, 1197–98, 1209, 1211, 1212, 1213, 1221, 1243, 1281, 1322, 1356
- Cambridge, 37, 41
- Camerer, C. F., 876
- Camilli, G., 119, 196, 236, 240, 404, 408, 410, 411, 412, 421
- Campbell, C., 1045
- Campbell, D. T., 201, 202, 221, 1160, 1357
- Campbell, H., 1055
- Campbell, N., 1521
- Campbell, N. R., 740
- Campbell Allen, A. J., 110n1
- Campbell's law, 1160–61
- Campion, M. C., 1382
- Cannata, M. A., 1137
- Cao, M., 1380–81
- Capshe, J. H., 59
- Card, D., 1341
- Career and Technical Education (CTE), 1128
- CARES. *See* Coronavirus Aid, Relief, and Economic Security
- Carless, D., 1018, 1026
- Carlin, B. P., 1456
- Carlson, D., 1037
- Carlson, J. E., 75, 76, 82
- Carlson, J. F., 1346
- Carnegie Corporation, 89
- Carnegie Foundation for the Advancement of Teaching, 76, 1219
- Carnegie Institute of Technology, 73
- Carnegie Mellon University, 623
- Carnevale, A. P., 423
- Carnoy, M., 140, 1144, 1425, 1484, 1485, 1488
- Carroll, J. P., 1350
- Carter, P. L., 1071
- Case, 363
- CASEL. *See* Collaborative for Academic, Social, and Emotional Learning
- Casner-Lotto, J., 1343–44
- CAST, 973
- Castellano, K. E., 793, 1152, 1154
- Castex, G., 1341
- CAT. *See* computerized adaptive testing
- categorical classification decisions, 337–38
- Categorical Model From Delaware's 2009–2010 School Year, 1151t
- categorical models, 1152–53, 1155
- Cattell, J. M., 51, 1520, 1522
- Cattell, R. B., 1377, 1387, 1388, 1389
- Catterall, J. S., 1120
- causal inferences, validity and, 218–19
- causal linkages, 1425
- Cawthon, S., 969
- CBAL. *See* Cognitively Based Assessment of, for, and as Learning
- CBM. *See* curriculum-based measurement
- CCR. *See* college-and-career-readiness
- CCSS. *See* Common Core State Standards
- CCSSO. *See* Council of Chief State School Officers
- CDM. *See* cognitive diagnostic model
- CEFR. *See* Common European Framework of Reference for Languages
- Cengage, 1344
- Census Bureau, U.S., 972, 980
- Center for Education Statistics, 1208
- Center for Measurement Justice, 11
- Center for Reinventing Public Education, 3–4
- Center on Education Policy, 1139
- Central Components of the Process of Measurement, 1536f
- A Century of Standardized Testing* (Linn), 384
- certification, 22, 1266–69; ADA relation to, 1283–84; assessment development for, 1277; employment tests and, 1273–74; feedback and, 1295; Golden Rule procedure and, 403; K–12 assessment and, 755–56; KSAs and, 214; licensure compared to, 1337n1; panelists for, 1314; professional development and, 1321–22; program sponsors for, 1274–75; regulation of, 1270–73; reliability of, 1297–98; scale scores in, 760; scaling in, 737; score reporting for, 936; speeded assessments for, 354; stakeholders and, 1313; TBA for, 586; teacher, 1281–83; Teaching Exercise and Fitness, 871; test security for, 1286–87; training for, 1278; validation of, 1322–23
- Certified Financial Planner, 1276
- Certified Public Accountants, 1293
- chained equating, 777–78, 779, 782, 791
- Chakrabarty, R., 552, 1144
- Chalhoub-Deville, M., 832
- Challman, 275n2
- Chan, D., 1382, 1383
- Chang, H. H., 1306, 1537, 1549
- Chang, T. H., 533
- Chapelle, C. A., 211, 212, 233, 275n4
- Chapman, L. H., 1060
- Chappuis, J., 1071
- character skills, 23, 962, 1340, 1345, 1346, 1355, 1356
- Charcot, J. M., 47–48
- Charness, G., 1387
- Chartered Financial Analyst Institute, 1276
- ChatGPT, 130
- Chattergoon, R., 998
- Chauncey, H., 87
- Chavira, D. A., 127
- cheating, 614; classification and, 616; item exposure and, 479; machine learning and, 520; process data and, 615; psychometrics and, 519; remotely administered exams relation to, 400, 402; teacher, 1148; test administration and, 562; test security and, 397, 518; topic models and, 542
- Chemudugunta, C., 542
- Chen, F. F., 707
- Chen, G., 217
- Chen, H., 539
- Chen, H. H., 779
- Chen, J., 850, 851
- Chen, L., 539, 541
- Chen, M., 556
- Chen, W. H., 672, 694, 706
- Chen, X., 1233, 1234
- Chen, Y., 549, 686
- Cheng, P. E., 781
- Cheng, Y., 313, 540
- Chernyshenko, O. S., 1351, 1377
- Cherry, K., 73
- Chetty, R., 1141, 1154, 1197, 1198
- Chi, M. T. H., 119
- Chin, M., 402
- China, 132; cheating in, 518; civil service examinations in, 15–16, 35, 37, 83, 384–85, 436n1; National College Entrance Examination in, 90
- Ching, C. C., 158
- Chinn, R., 1287, 1288, 1289–90
- Chiorri, C., 415
- Cho, Y., 233
- Chodorow, M., 534, 535, 538
- Choi, H.-J., 547, 548, 659, 668, 669, 701
- Choi, J., 357, 547
- Chojnacki, G., 1035
- Christensen, L. L., 981
- Christian, M. S., 1382, 1383
- Christiansen, N. D., 1381
- Christman, J. B., 1034–35, 1036
- Chrysafiadi, K., 626, 629
- Chudowsky, N., 1138, 1140
- Chung, G. K., 699, 733n1
- Churchill, Winston, 56
- Circi, R., 1039, 1064, 1068, 1069–70
- circles of interest, 921, 922f, 922–23
- Cisco learning network, 500
- Citro, C.
- civil life, test scores relation to, 129
- civil rights, in ADA, 142
- Civil Rights Act (1964), 139, 964, 1269, 1280, 1282
- Civil Rights Act (1991), 422
- civil rights movement, 9, 34–35
- civil service examinations, in China, 15–16, 35, 37, 83, 384–85, 436n1

- Cizek, G. J., 206, 218, 221, 518, 520, 822, 828, 852, 853, 856, 858, 859, 860, 861, 863, 877, 1022, 1288, 1307, 1310, 1312, 1313, 1316, 1547, 1551
- Claims and Evidence in Principled Assessment Design, 487*t*
- Clark, C. L., 60, 1292
- Clark, D., 1341
- Clarke, M., 1431
- Class Action Lawsuit*, 401
- classical congeneric model, 295
- Classical Item Statistics, 704*t*
- classical test theory (CTT), 15, 36, 37, 93, 1525; binomial error model relation to, 327; correlation and, 41; DCM compared to, 348; dominance response process and, 1377; equating and, 778; equity and, 768; error variance and, 295; forced-choice methods and, 1380; generalization inferences in, 1302; GT and, 63, 111*n*18, 297, 298, 1527, 1528; IRT relation to, 74, 602, 1529, 1530, 1533–34; linear error model and, 1520; as mathematical or statistical model, 1524; MST and, 351; parallelism in, 296, 300; Rasch model and, 1533; reliability and, 64–66, 278, 280–81, 286, 412, 1359–60, 1361; reliability coefficients and, 285, 288, 302–3, 311; RSM and, 1367; SEM in, 289–90, 310, 324; test scores and, 67, 1529; true score and, 281–82, 286–87, 318, 1525, 1526, 1531; unidimensional pairwise preference task data and, 1378; variance in, 302–3, 325
- classification, 616, 1563
- classification accuracy, 23, 226, 347–48, 1317–18; classification consistency compared to, 285; reliability and, 1269; true score and, 338–39, 342
- classification category, 284
- classification consistency, 338–39; classification accuracy compared to, 285; cut score and, 285
- classification indices: IRT and, 345; raw-to-scale score conversions for, 347
- class modeling, fairness and, 164
- classroom assessments: data collection from, 657–58; instructional decisions relation to, 1050; large-scale testing and, 118, 252, 253; PAD and, 511*n*1; reliability of, 1300
- classroom formative assessments, 20, 1022, 1024*t*, 1028–30, 1070, 1078–79; CBM compared to, 1056; data collection and, 1036; feedback for, 1023, 1025, 1031–32
- classroom observations, 424, 1142
- classroom questions, 1027–28
- Clauser, A. L., 12, 90, 1288
- Clauser, B. E., 12, 90, 110, 110*n*6, 132, 255–56, 258–59, 586, 596–97, 599, 619, 803, 841, 851, 867–68, 870, 871, 873, 880, 897, 1201, 1270, 1281, 1288, 1297, 1300–1301, 1312, 1317, 1318, 1520
- Clauser, J. C., 12, 586, 596–97, 599, 619, 868, 870, 873, 1314
- Claypool, T. R., 993
- Clayton, B., 45
- Cleary, T. A., 82, 388, 416–17, 421, 518
- Clemans, W. V., 1377
- CLEP. *See* College-Level Examination Program
- Clevenger, J., 1382, 1386
- click-stream data, 942
- Cliff, N., 1553
- Clifford, G. J., 73
- Cliffordson, C., 1043
- Clifton Strengths, 1355
- Clinedinst, M., 1209
- CLR. *See* cultural and linguistic responsiveness
- Clune, W. H., 1021, 1033
- cluster analysis, 255
- coaching: for admissions testing, 396–97; bias in, 1162; test scores relation to, 1191
- Coburn, C. E., 228
- Cochrane, R. C., 1342
- Cochran-Smith, M., 1071
- The Code of Fair Testing Practices in Education*, 397, 400
- Coding Speed test, 1387
- Coe, M., 1031
- coefficient alpha, 36, 66; GT and, 381*n*1; number-correct scores and, 280; reliability coefficients and, 297, 349, 361; reliability of, 294; RG and, 355; stratified, 297
- coefficient of equivalence, 290
- coefficient of stability and equivalence, 290
- Coetzer, E. P., 1342
- Cognia & Multi-State Alternate Assessment Consortium, 217
- cognition: assessment design relation to, 456–57, 458, 460; domain analysis and, 459; PAD and, 475; PLDs and, 457, 458–59; RPLDs and, 461, 463–64, 480; simulation-enabled performance tasks and, 596–97; TBA and, 590; TEIs and, 592
- cognition model, 151, 152, 153*f*, 153–54, 167
- cognitive ability, 1341, 1343, 1362, 1368–69, 1381–83; employment tests and, 962; grit compared to, 1347; learning outcomes and, 1238; noncognitive skills relation to, 1390–91; SJTs and, 1386; TIMSS and, 1438; validity and, 1043
- cognitive activity, 119, 123, 131
- cognitive and social model, of standard setting, 828, 830*f*, 833–34, 847–48, 873–74, 879
- cognitive complexity, 457, 470*f*, 471*f*; alignment and, 1160; borderline test takers and, 842–43; PAD and, 469, 470; standard-setting design and, 876
- cognitive data summary tables, 1463
- cognitive demand, 156, 457, 465, 466; difficulty drivers and, 467; item type relation to, 473
- cognitive design systems, 445
- cognitive diagnostic model (CDM), 627, 628, 1075; reliability and, 1298; score reporting and, 943
- cognitive domain model, 635–36, 637
- cognitive evidence, 251, 1066
- cognitive feedback, 1025
- cognitive-judgmental task, 18–19, 848; bias in, 846; in Bookmark method, 867–68, 873; borderline test takers relation to, 869; for decision-making, 843–44; individual panelist cognitive processing submodel and, 831; item difficulty and, 850; panelists and, 833–34, 841, 852, 865, 876–77; PLDs and, 829, 834, 835*t*, 837; standard-setting design relation to, 857, 861; threshold-signal detection judgments and, 840; training for, 860–61; value judgments and, 840; in workshop process submodel, 856
- cognitive load, 842
- Cognitively Based Assessment of, for, and as Learning (CBAL), 219–20
- cognitive models, in large-scale testing, 475
- cognitive process, 553; for classroom questions, 1027–28; in standard setting, 841; TBA and, 611
- cognitive psychology, 91
- cognitive scales, 1457–58
- cognitive science, 897, 929, 945
- cognitive skills, 23, 1340–41, 1390–91, 1478; educational attainment relation to, 1434, 1469, 1471; IQ and, 1345; scaling and, 1479
- cognitive-social model, 834
- cognitive strain, 842, 845
- cognitive tasks, 829
- cognitive theory, 551–52, 628
- Cohen, A. S., 17, 515, 517, 897
- Cohen, D. K., 1146
- Cohen, J., 339, 357, 1143, 1348, 1367
- Cohen, S., 163, 164
- Coh-Metrix, 545, 547
- Cole, M., 116, 126, 130, 235, 388, 964, 1191
- Cole, N. S., 408, 421–23
- Coleman, J. S., 977, 979, 1119
- collaboration, 245, 994
- Collaborative for Academic, Social, and Emotional Learning (CASEL), 1340, 1345, 1348, 1350, 1355, 1356–57
- collaborative problem solving assessments, 259–60
- college-and-career-readiness (CCR), 1128
- College Basic Academic Examination, 870
- College Board, 36, 385, 528, 744, 788, 789, 1202, 1203–4, 1210, 1211, 1212, 1226–27, 1230, 1236–37
- College Entrance Examination Board, 75, 87, 385, 1201–2
- College-Level Examination Program (CLEP), 766

- College Placement Tests (CPTs), 585. *See also* placement tests
- college-readiness tests, 143
- Collegiate Learning Assessment, 1239
- Collier, V. P., 974
- Collins, A., 460
- Collins, J., 130
- Collis, J. M., 1359
- Collis, K. F., 459
- Columbia University, 10, 26, 51, 53, 73, 76
- Comande, G., 622
- Commission on the Future of Higher Education, 1239–40
- Committee on Methods of Psychological Examining of Recruits, APA, 53
- Common Core Standards Writing Team, 1063
- Common Core State Standards (CCSS), 442–43, 533–34, 556; alignment and, 1158, 1159; AP and, 446; cognitive complexity for, 469, 470; domain analysis and, 461, 463; Mathematics, 455; SAT relation to, 1203; TBA and, 587
- Common European Framework of Reference for Languages (CEFR), 534, 538, 555, 579n1, 871, 894n11
- common factor model, 663
- common formative assessments, 20, 1038; assessment design of, 1040–41; instructional planning relation to, 1039
- common-item assumptions, 605, 606
- common-item equating, 605, 606–7, 771
- communication, score reporting as, 901–2, 917
- Community Training and Assistance Center (CTAC), 1060
- comparability, 762; bias relation to, 605; CLR relation to, 242; consequential tests and, 608; COVID-19 pandemic and, 795; fairness and, 196; ILSAs and, 1436–37, 1441, 1448, 1472, 1477, 1481; item parameter invariance and, 1510n51; linking and, 786; PISA and, 1485; psychometrics and, 607; score, 606; of summative assessments, 658; TBA and, 610, 634–35; validity and, 194–95
- comparative judgment, law of, 1518–19
- Comparison of English-Based Automated Essay Scoring (AES) Systems, 535t
- Comparison of Non-English-Based Automated Essay Scoring (AES) Systems, 559t
- Compass and Asset, 1234
- compensatory education programs, 56
- competency-based education, 23, 150, 207
- complex assessment types, IRT and, 338
- complex scoring approaches, 749
- composite scaling, 310, 320–22, 335, 420, 1437
- composite scores, 284, 320–21, 346; ACT, 1206, 1208; battery scaling and, 757, 792; error variance for, 296, 306; reliability of, 361; as scale scores, 322; UIRT and, 348
- composition-rhetoric theory, 553
- compound binomial model, 344
- Compton, D. L., 1055
- computational psychometrics, 34, 1075, 1320
- computer-assisted instruction (CAI), 584–85, 622–23, 624
- computer-assisted testing, 91
- computer-based assessment: accessibility and, 495, 497; cheating and, 519; data collection in, 1308; ILSAs as, 1439; interim assessment as, 1034; in K–12 assessment, 498, 1073; large-scale testing and, 515–16; LSAT, 1221; MC in, 529; PIAAC as, 1431, 1434; PISA as, 1429–30; TEIs and, 590–91; test accommodations for, 517; test security for, 1286; TIMSS as, 1427; TOEFL iBT, 534; UX and, 498–99
- computer-based performance tasks, 456
- computerized adaptive testing (CAT), 61, 366, 586, 676, 1072; accessibility and, 972; data collection in, 1074; GRE and, 585, 1213; for interim assessments, 1073; intraclass correlations and, 352; IRT and, 82; linking and, 794, 795, 798; proficiency estimates in, 351, 353; psychometrics and, 584; reliability of, 367; speeded assessments and, 354; for summative assessments, 658; test security relation to, 1213
- computer scale-aligning methods, 777
- computer science, 656
- computer skills, 530
- Conant, J. B., 87
- Conaway, C., 1470
- Concentric Circles of Interest in Assessment Results, 922f
- conceptualization, 279–80
- conceptualization-estimation framework, 301
- Conchie, B., 1355
- concordances, 786, 788–89, 1205
- concurrent calibration, 605
- conditional distribution, 663
- conditional error variance, 289, 312–13, 323
- conditional independence, 78, 662, 669–70, 674, 680; conditional RP and, 683, 685; IRT and, 671, 1530; MIRT and, 707
- conditional indices, 346
- conditional inference, 124–25, 165, 166–67
- Conditional Inference About Student Knowledge and Abilities, 166f
- conditional linking, 797
- conditional probability selection model, 422
- conditional probability tables (CPT), 629
- conditional reasoning, 1388
- conditional response probability, 682, 683, 685
- conditional scale-score variance, 323
- conditional SEM (CSEM), 284–85, 320, 324, 366; binomial error model, 326–27; conditional error variance, 289; CTT and, 286; in GT, 328–29; multivariate GT and, 334; nonlinear transformation of, 329–30; raw scores and, 331–32, 335; reliability and, 291, 362, 367; reliability coefficients and, 324–25; scale scores and, 289, 330, 331, 332, 333, 335, 336–37; stakeholders and, 1299; true score and, 325, 331; in UIRT, 333
- conditional standard errors of measurement (CSEMs), 753–54, 756, 760, 1297
- conditional variance, of MLE, 329
- conditioned latent regression model, 680
- conditioning regression model, 684
- Condon, D. M., 361, 1356
- confidence intervals, 366, 1124, 1295; reliability and, 349, 357, 360; SEM and, 324; true score and, 750–51
- confirmation bias, 209, 216, 846
- confirmatory factor analysis, 349
- Confrey, J., 1063, 1064, 1066, 1068–69, 1075
- Confucius, 83
- congeneric forms, 291–92, 295–96
- Conger, A. J., 1367
- Connelly, B. S., 1368, 1369, 1370
- Connolly, J. A., 613
- consequences: of accountability policies, 1117; of assessment design, 228; of decision-making, 229; decision/use inferences and, 249, 259; fairness of, 964; of grade promotion exam, 1141; of language testing programs, 232, 233–34; of MCT, 1122; of performance assessment, 230–31; of performance standards, 224–25; proximal, 437n3; of score interpretation, 232, 233; of test-optional policies, 425; test use and, 206, 221–22, 224, 228–29, 232, 233, 234, 235, 260; unintended, 220, 1147–48, 1155; validation and, 232, 234; validation framework and, 221–22, 224; validity argument and, 229–30, 389; validity relation to, 193, 222–24, 275n3
- consequential evidence, validation framework and, 197
- consequential tests: automated scoring in, 653n2; comparability and, 608; cut score and, 754; TBA and, 584–89, 600, 607, 631, 633–34, 636–37
- consequential validity, 724, 1476, 1488; growth model analyses for, 230; of ILSAs, 1484–85, 1491
- consistency, of reliability coefficients, 360
- constant ratio model, 421
- Constitution, U.S., 1280
- construct-based inferences, 611
- construct-centered measurement, 445
- construct definition, 118, 217, 622, 982; AP and, 446–47; assessment design and, 258; in rubrics, 620, 621
- constructed-response items (CR): automated scoring of, 584; in large-scale testing, 245; linking and, 795–96; RP values in, 759; score weights and, 743–44; test scores for, 1454

- construct fidelity, 590, 592; SBTs and, 596; TE and, 984
- construct-irrelevant variance, 156;
accessibility relation to, 965; assessment design and, 454–55; assessment performance relation to, 391–92; bias and, 1308; CLR and, 497–98; in criterion setting, 259; DIF relation to, 413, 990; disabilities and, 166, 399; diversity relation to, 468; fairness and, 156, 218, 966, 997; in GRE Analytic Writing Assessment, 256–57; ILSAs and, 1481; KSAs and, 258; in performance assessment, 1388–89; PLDs and, 458; polarizing topics as, 395–96; rater language background and, 237; RPLDs and, 478; score interpretation relation to, 517, 562; SES as, 980; TBA and, 550, 601, 1320; TE and, 984; test accommodations and, 391, 516–17, 960; in test administration, 398; test-based inflation and, 1160, 1162–63; test scores and, 207, 394, 399; UDA relation to, 973; validity and, 227, 282, 597–98
- Construct Map Showing a Learning Progression for Student Understanding of Earth in the Solar System, 1065t
- construct-relevant variance, 137, 155–56; CLR and, 497–98; in digital assessments, 1483; human raters and, 620; PLDs and, 458; validity and, 610–11
- construct representation, 203; in collaborative problem solving assessments, 260; in digital assessments, 1482; in large-scale testing, 130; “theory of the construct” and, 245
- construct underrepresentation, 207, 394, 551
- construct validity, 91, 198, 200–204, 208; ILSAs and, 1474; multilevel, 217–18; test interpretation and, 205, 206
- content-based standard setting, 1315
- content expertise, 870
- content model, of validity, 199
- content referenced growth, large-scale testing and, 253
- content-related evidence, validation and, 393
- content standards, 976, 992, 1048–49, 1056, 1310, 1546; accountability in, 227–28; ESSA and, 970, 1155–56; NCLB and, 133; panelists and, 833, 867–68; performance standards relation to, 1121–22, 1310; professional development and, 1040; student achievement and, 1042; summative assessments and, 658. *See also* standard setting
- content validity, 585, 1237; achievement tests and, 200–201; alignment and, 489; construct validity and, 202
- contextual information, for score reporting, 921
- contextual material, 395–96
- Continuing Board Certification, 1307, 1309, 1319
- The Continuing Board Certification*, 1321–22
- continuous testing, 1307–8
- Contrasting Groups method, 841
- A Contribution to the Study of the Subintestinal Nervous System of Insects* (Binet), 48
- Control Data Corporation, 584–85
- conventional path diagrams, 713
- Conventional Test Specifications With Separate Considerations of Components, 475f
- Cook, L. L., 408, 968, 969, 1191–92
- Cook, R., 477–78, 480, 722, 877–78, 1547
- Coombs, C. H., 740, 741
- Cooper, H., 1390
- Cooperative Achievement Tests, 765
- Cooperative Test Service, American Council on Education, 59
- Cope, B., 946
- Copyright Act (1976), 1280
- Coraggio, J. T., 849
- Corbet, H. D., 148
- Corbett, A. T., 623, 628, 630
- Corcoran, S., 1155
- Corcoran, T., 1062
- Cordray, D., 1037
- CORE. *See* California Office to Reform Education
- Cornfield, J., 69
- Coronavirus Aid, Relief, and Economic Security (CARES), 6
- correlation, 38, 40–41, 42, 112n20; partial, 43–44; prediction and, 91
- correlational evidence, as validity, 199
- correlational psychology, 37, 42, 43, 63
- correlational science, 44
- “Correlation Calculated From Faulty Data” (Spearman), 44
- correlation coefficient, 40–41, 42–44, 46, 92, 1559; CTT and, 286; GPA and, 1230; intraclass, 311–12, 317; validity and, 233
- Corrin, L., 25, 928
- Corstjens, J., 1383–84
- Cortellessa, E., 25
- Costa, A. L., 1356
- Costa, P. T., 1342
- Cotter, R. A., 1220
- Cotton, K., 1027–28
- Cottrell, J. M., 977
- Coughlin, E., 1034
- Council of Chief State School Officers (CCSSO), 442, 446, 448, 981, 1020, 1047, 1121
- Council of Europe, 538
- Council of Graduate Schools, 1214, 1215
- Courtney, M. G. R., 1470
- Couture, J.-C., 1484
- covariance/correlation structure, 669, 676
- COVID-19 pandemic, 4, 29n1, 1220; accessibility and, 408; accountability and, 4–5; accountability testing and, 385, 1114; admissions testing and, 15, 22, 400, 1191, 1212, 1213; alternative online administrations and, 16; ESSA relation to, 1130; graduate school admissions relation to, 1213; LSAT relation to, 1221; mental health and, 3; remote learning and, 136; remotely administered exams and, 15–16, 384, 400; remote proctoring and, 634; social justice relation to, 1192; survey congestion and, 1480; SWDs and, 407; TBA and, 589, 608; test administration and, 795, 803; test-based accountability and, 1169; test-optional policies and, 21
- Covington, M. V., 1042, 1046
- Cowell, W. R., 765
- Cowen, J., 606
- Cowen, J. M., 1144
- Cox, C. B., 1382
- CPT. *See* conditional probability tables
- CPTs. *See* College Placement Tests
- CR. *See* constructed-response items
- Craig, A. T., 436n2
- Cravens, 868, 870, 874
- creation science, 7
- creativity, 1238, 1345, 1350, 1353, 1355, 1366–67
- Credé, M., 1347
- credentialing, 22, 60, 193, 795, 1266, 1302; assessment development for, 1287–89; cut score in, 1318; expert panels for, 854; extrapolation for, 226–27; fairness in, 260–61, 999, 1269; feedback for, 1296; games and simulations for, 1294; generalization inferences and, 1303; group reports and, 901; innovative item types for, 1292–93; mandatory professional, 1275–76; panelists and, 1315; performance in practice relation to, 1304; performance standards for, 822; practice analysis and, 1289–91; program sponsors for, 1274; in psychology, 1267; reliability in, 23, 1298–1300; score reporting for, 922; stakeholders and, 1277–78, 1314; standard setting for, 824, 833, 853, 859, 879, 1270, 1309–12, 1317; test accommodations, 1280; validity in, 23, 1269, 1273, 1303, 1309, 1323; voluntary professional, 1276–77
- credit-bearing courses, 1233–34
- Crenshaw, K. W., 9, 986
- Crisp, V., 194, 219
- Criterion Online Writing Evaluation service, 534, 536
- criterion-referenced feedback, 861, 1036
- criterion-referenced tests, 338, 1118–20, 1210; reliability for, 1298; score reporting and, 62, 898, 919
- criterion-related validity, 449, 556
- criterion setting, construct-irrelevant variance in, 259
- criterion validity, 199–201

- critical incident technique, 59, 60
critical interpretation, of test scores, 131
critical race theory, 8–9, 12, 29n2
Croft, M., 88
Cromley, J. G., 974
Cronbach, L. J., 13, 14, 34–35, 36, 57, 64, 66–67, 69–70, 90–91, 92, 111, 111nn17–18, 112, 112n19, 125, 163, 192, 196, 197–99, 201–4, 205, 206, 208, 210, 213–14, 215, 218, 221–22, 226, 227, 237, 279, 294, 297–98, 304, 353–54, 360, 366, 368, 384, 389, 739, 1119–20, 1313, 1359, 1367, 1372, 1526, 1528, 1550, 1554
Cronbach's alpha, 1526
Cronin, J., 1148
Crook, A. E., 1382
Crooks, A. D., 1044
Crooks, T. J., 1020, 1046
Cross, L. H., 1316
cross-classified data structure, 662, 685
cross-classified IRT, 716, 717, 717t, 718, 721t; multidimensional IRT *versus*, 720, 720t, 721; RD and, 722–23
Cross-Classified Item Response Theory Model: Data Structure (Example 2), 720
Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Data Structure, 717t
Cross-Classified Item Response Theory Model With Individuals and Items as Crossed Factors: Key Results, 717t
cross-cultural equivalence, 1456
cross-cultural intelligence testing, 155
Cross-Sectional Comparisons of CORE District Students Scale Scores Across Social-Emotional Learning (SEL) Scales and Achievement Test Scores, 1364f
Croudace, T. J., 1380
Crouse, K., 1061
crowdsourced learning, 927–28
Cruce, T. M., 1197, 1206, 1208
Csapo, B., 590, 597
CSEM. *See* conditional SEM
CSEMs. *See* conditional standard errors of measurement
CTAC. *See* Community Training and Assistance Center
CTB/McGraw-Hill, 86, 789
CTE. *See* Career and Technical Education
CTT. *See* classical test theory
cubic spline postsmoothing, 774
Cudeck, R., 676–77
Cui, Y., 1298
Cui, Z., 774
Culpepper, S. A., 418
cultural and linguistic responsiveness (CLR), 20, 242, 256, 994; accessibility through, 963, 966, 968, 978–79, 988–90; in assessment design, 235, 258, 494–95, 990, 991; assessment development and, 997, 999; fairness and, 497–98, 966, 989
cultural diversity, 122
culturally responsive assessment, 243–44, 245–46, 247t, 248t; accountability relation to, 1168–69; equity relation to, 235–36, 237, 239
culturally sustaining assessment, 1167–68
cultural relevance, of assessment tasks, 121
Cultural Relevance Versus Construct Relevance, 437n5
cultural validity, 992–93
culture, 138–39, 162, 396, 986; assessment design relation to, 235, 242–43; assessment performance relation to, 236–37; language and, 13–14, 116, 117, 119–20, 121–22, 975; learning outcomes relation to, 126–27; learning progressions relation to, 156; licensure relation to, 261; NCLB relation to, 140; PAD relation to, 446; societal needs relation to, 129; UDL relation to, 245
culture-fair selection, 421–22
cumulative logits, graded response model with, 682–83
Cunningham, S., 1475
Cureton, E. E., 91, 198, 200, 1528–29
Curran, P. J., 701
curricular validity, 150–51
curriculum-based measurement (CBM), 20, 1052–53, 1054f, 1055; data collection and, 1056, 1058; instructional decisions and, 1059; state summative assessments and, 1054; TOA for, 1057
Currier v. NBME (2012), 1285
Cushing, E., 1061
cut score, 18, 338, 754, 760, 870, 893n6; admissions testing and, 1199; assessment design and, 480; in categorical models, 1153; classification accuracy and, 339; classification consistency and, 285; on criterion-referenced tests, 1119–20; for decision-making, 225–26; fairness and, 421, 422; false positives and, 1317–18; Framework to Guide Workshop Design and Gather Validation Evidence to Support Claims about the Interpretation of, 857t; in high-stakes testing, 1297; in HSEEs, 1138–39; impact data relation to, 861–64; in large-scale testing, 347; for MCT, 1120; NCLB and, 866; performance standards and, 474, 832; primary scales and, 760; in principled item development process, 877–78; reliability and, 363; RPLDs and, 477; standard-setting method and, 224–25, 859; standard-setting results and, 863; test performance relation to, 756; value judgments and, 840; in workshop process submodel, 856
cut scores, achievement levels and, 927, 1556, 1557
Dadey, N., 250, 251, 1021, 1033, 1034, 1051
D'Agostino, J. V., 1157
Dalessandro, S. P., 1230
Daly, J. A., 847
Damian, R. I., 1346
Daniels, L. M., 944
Danish Institute for Educational Research, 81
Darling-Hammond, L., 118–19, 1028, 1047, 1071, 1482
Darlington, R. B., 422
Darr, W., 1375
Darwin, C., 7, 34, 37–39, 92
dashboard, 923–25, 928, 1133
data-based individualization (DBI), 1058
Data-Based Program Modification, 1056–57
data cleaning, 1460, 1461
data collection, 540, 603, 681, 794, 1074, 1359; for assessment design, 775–76; for assessment development, 1290–91; CBM and, 1056, 1058; from classroom assessments, 657–58; in digital assessments, 1459–60; equating and, 738t, 770, 773; ILSAs and, 1424–25, 1439, 1449, 1466, 1471, 1474–75, 1488; information technology for, 1076–77; interim assessment and, 1034–36, 1037–38; large-scale sampled surveys and, 659; latent variable models and, 679; linking and, 738t, 800, 803; for machine scoring, 526–27; NAEP and, 679, 1354; NEAT design for, 771; psychometrics and, 1533; from PUF, 1463; for reliability, 290; in response process, 612–13; for response time, 1307, 1308; software-based, 724; state summative assessments and, 1048–49; with surveys, 934–35; TBA and, 660; test specifications and, 1289, 1290
data confidentiality, 1466
Data Explorer, NAEP, 923, 1464–65
data logging, 254–55
data mining, 254, 617, 1075, 1391
data quality, 1136, 1438–39, 1445, 1449, 1460, 1465
The Data Quality Campaign, 1133
data security, 1460, 1465
data smoothing, 773–74, 781
data story, 911–12, 913, 913f, 914
data use TOA, 1034–35
data validity, 527
data visualization, 19, 897, 924, 945–46
Davenport, E. K., 1200
Davey, T., 367, 772, 776, 799
Davidson, E., 1123, 1124, 1125
Davidson, K. L., 1033
Davies, M., 1239
Davis, B., 1023
Davis, L., 1243
Davis-Becker, S., 18, 846–47, 853, 860, 868, 873, 877, 1311, 1313–14
Davison, M., 802
Dawber, T., 867–68
Dawes, R. M., 841
Day, D. V., 524

- DBI. *See* data-based individualization
- DCM. *See* diagnostic classification models
- D'Costa, A. G., 1290
- Dean, P. R., 45
- Deane, P., 594, 595, 614, 617, 1062
- Deary, I. J., 1034
- de Ayala, R. J., 163, 164, 1361, 1421n4
- Debeer, D., 1388
- Debnath, S. C., 1375
- De Boeck, P. D., 681, 723, 927–28, 1535, 1551–52
- Debra P. v. Turlington*, 150, 388, 397–98
- DeCarlo, L. T., 357
- de Carvalho, A., 616
- Dechter, E. K., 1341
- decision-making, 20, 150, 233, 1136; BEAR Assessment System and, 1069; bias in, 147, 841–42, 846; cognitive-judgmental task for, 843–44; consequences of, 229; cut score for, 225–26; fairness and, 963; grading and, 1043; in individual panelist cognitive processing submodel, 833; interim assessment relation to, 443; international assessments and, 140; large-scale testing for, 133–34; milestone ratings and, 1321; PAD for, 445; of panelists, 861; psychology relation to, 834; reliability and, 367; sociocultural context for, 151, 152; stakeholders and, 964; standard setting and, 822, 824; TBA and, 17, 582, 622, 630, 632–33; test scores and, 215, 1508n27; validity and, 1550–51
- decision models, 617
- decision studies (D studies), 298, 300, 301, 303–4, 305–6, 328
- decision/use inferences, 211, 233–34, 257, 275n4; consequences and, 249, 259; validity and, 215
- Decristan, J., 1031
- Dee, T. S., 990, 1138–39, 1143–45, 1147, 1148, 1168
- deep knowledge tracing (DKT), 628
- Deerwester, S., 542, 544, 555
- deficit interpretations, 1117
- Definition and Selection of Competencies project, 1434
- De Fruyt, F., 1348, 1350, 1351
- de Klerk, S., 1320
- Delandshere, 1473, 1474
- de la Torre, J., 669, 1296
- Delaware Department of Education, 1153
- De Leeuw, J., 664
- De Leng, W. E., 1384
- Delgado, C., 156
- Delton, J., 935
- DeLuca, C., 1029
- DeMars, C. E., 320, 498, 1240
- DeMauro, G. E., 442, 456
- DeMers, S. T., 1267
- Deming, D. J., 1122, 1146, 1344
- Dempster, A. P., 686, 1534
- Dendir, S., 516
- DeNisi, A., 1023, 1025
- Deno, S. L., 1053, 1054, 1055, 1056–57
- Dental Interactive Simulation Corporation, 1319
- Denver, B., 1063
- Department of Defense, U.S. (DoD), 58, 60–61
- Department of Education, U.S. (DoE), 26, 89, 386, 443–44, 483, 500, 602, 611, 879, 927, 971, 972, 1041, 1049, 1051, 1077, 1130, 1151, 1152, 1190–91, 1208, 1209, 1211, 1239–40, 1507n12
- Department of Government Efficiency, U.S. (DOGE), 25
- Department of Health and Human Services, U.S., 26
- Department of Homeland Security, U.S., 26–27
- Department of Justice, U.S., 145
- Department of Labor, U.S., 980
- DePascale, C. A., 20, 118, 124, 194, 242, 249, 386, 602, 608, 622, 657, 926, 962, 1061, 1115
- dependability coefficient, 303, 722
- The Dependability of Behavioral Measurements* (Cronbach), 66, 69–70, 298
- Dept, S., 1436, 1507n17
- Depth of Knowledge (DOK), 459, 474, 474
- derived accountability scores, 19
- Descriptive Statistics of Latent Pretest Score and Latent Change Score by Condition, 715t
- Descriptive Statistics of Pretest and Posttest Sum Scores by Schools and Conditions, 702f
- Descriptive Statistics of Sum Scores of Pretest and Posttest by Condition, 701t
- design-based inference, large-scale sampled surveys and, 659
- Design of Experiments* (Fisher), 47
- Design Pattern Organization and Grain Size, 466f
- design patterns: in PAD, 465, 466, 467; RPLDs and, 467–68
- design science, 451
- Desmarais, M. C., 626
- Desruisseaux, T. M., 1167
- deterministic-input noisy “and” gate (DINA), 668, 703, 709, 710
- Deunk, M. I., 871, 874
- Derville, C., 832
- Devlin, J., 549, 555
- Dewar, J., 39
- Deygers, B., 234
- Dhuey, E., 1474
- diagnostic classification models (DCM), 348, 710; IRT compared to, 349–51, 668; item analysis and, 703; with item factor models, 669; for K–12 education, 724; latent variable models and, 683, 709; MIRT compared to, 668, 711; Q-matrix for, 709, 710t; Top Five Estimated Proportions of Attribute Profiles in, 711t
- diagnostic feedback, 535
- diagnostic scoring, in large-scale testing, 1018
- diagnostic testing: ELs and, 147; fairness and, 146; K–12 assessment and, 1027; score reporting for, 917; special needs and, 144
- Diamantopoulos, A., 663
- Diaz, J., 26
- Diaz-Bilello, E., 1060, 1061, 1062
- DiCerbo, K., 137, 493, 590
- dichotomous items, 410–11, 1529–30
- Dickson, J. H., 39
- Dickson-Markman, F., 847
- Dictionary of Occupational Titles, 1344
- Dietel, R., 1164–65
- DIF. *See* differential item functioning
- differences-that-matter criteria, 780
- differential-boost hypothesis, 969–70
- differential item functioning (DIF), 404, 424, 437n7, 1370; assessment development and, 414, 988; BARS relation to, 1375; comparability and, 194–95; construct-irrelevant variance relation to, 413, 990; for dichotomous items, 410–11; diversity relation to, 164, 235; fairness and, 119, 122–23, 423, 985–86; gender and, 987; IRT and, 163–64, 411, 438n8, 679; justice-oriented antiracist framework with, 996; PISA and, 240; psychometrics and, 479; standardization and, 412; test content and, 413–14
- differential prediction, 391, 419, 423, 424, 438n11
- differential test outcomes, for marginalized students, 228–29
- differential validity, 419, 423
- difficulty drivers, 465, 466, 467, 473
- Digest of Education Statistics, 401
- Diggs, C. R., 250, 251, 1021, 1033, 1034
- Di Giovinazzo, V., 1390
- digital assessments, 1456; assessment design of, 1440, 1450–51; construct-relevant variance in, 1483; construct representation in, 1482; data collection in, 1459–60; educational policy for, 1449; error control and, 1445; ILSAs as, 1447, 1455, 1477–78, 1483, 1489–90; sociocultural context and, 136. *See also* computer-based assessment; computerized adaptive testing
- digital delivery, 588, 599, 601, 608, 974–75, 1206
- digital environments, 137
- digital learning environments, 946, 1566
- digital literacy, 166
- digital platforms, 1436, 1439–42, 1445, 1476–77, 1483, 1490; ILSAs and, 1426; test administration on, 154, 158
- digital technologies, 1242, 1426, 1438
- Digman, J. M., 1342
- Dillon, G. F., 596

- dimensionality analysis, 666; in IRT, 697–98;
item analysis and, 702, 703; in item factor
models, 667; model fit for, 676
- dimensionality evaluation, 657, 702
- dimensionality reduction, LSA, 544f
- Dimitrov, D., 362
- DINA. *See* deterministic-input noisy “and”
gate
- Direct Consensus method, 827, 836, 838
- direct observation, 936
- disabilities: construct-irrelevant variance and,
166, 399; ESEA and, 62; fairness and, 123,
398; flagging policies and, 405–6; intellec-
tual, 48–49, 56, 147, 219–20; NCLB and,
63; test accommodations and, 135, 139,
141, 142–43, 167, 401–2; UDL and, 244.
See also Americans With Disabilities Act;
students with disabilities
- discipline specific words, 548
- discrimination, 26, 968; ADA relation to, 139,
142; bias and, 141; flagging policies as,
405–6; Fourteenth Amendment and, 149;
UC and, 406
- disparate impact, 150
- Distinguishing Characteristics of Principled
Assessment Design, 446f
- District of Columbia Public Schools, 1340
- diverse groups, 16, 192, 442
- diversity, 25; admissions testing and, 1243;
assessment design relation to, 149, 473–74;
CLR and, 994; construct-irrelevant vari-
ance relation to, 468; cultural, 122; cul-
turally responsive assessment relation to,
235–36; DIF relation to, 164, 235; fairness
and, 14–15, 135; in group interaction sub-
model, 831–32; large-scale testing and, 243;
panel, 852–54; personality tests and, 206;
social justice relation to, 245; test accommo-
dations for, 168; test-based accountability
and, 226; validity relation to, 194, 217
- Divgi, D. R., 351
- Division of Applied Psychology, Carnegie
Institute of Technology, 73
- Dixon-Román, E. J., 656
- Dizon-Ross, E., 1144
- DKT. *See* deep knowledge tracing
- D method. *See* quadrature distributions
- Dobbins, J. E., 1044
- Dobbins, M., 1486
- doc2vec, 549
- documentation, 365, 419, 897, 992, 1449,
1462; assessment development and, 485–
86; consequences and, 228; PAD and,
447, 451–52, 489, 497, 910; RG and, 355;
of standard setting, 1319; for test security,
1307; validity and, 210, 217, 483
- document-by-word matrices, 542f, 543, 544f
- DoD. *See* Department of Defense, U.S.
- DoE. *See* Department of Education, U.S.
- Doe v. National Board of Medical Examiners*
(1999), 406
- DOGE. *See* Department of Government Effi-
ciency, U.S.
- Doignon, J.-P., 623
- DOK. *See* Depth of Knowledge
- Dolan, R. P., 244, 984
- domain analysis, 454–55; CCSS and, 461,
463; cognition and, 459; PLDs and, 460;
task features in, 467
- domain definitions, 20, 790, 966, 982, 984,
992
- domain modeling, 452–53, 456, 478
- Domaleski, C., 231, 1139, 1140, 1211
- dominance response process, 1377
- dominant profile method, 843, 844
- Domingue, B., 755, 791, 990, 1553, 1561
- Donaldson, M. L., 1141, 1142
- Donlon, T., 1191, 1201
- Donnon, T., 1229
- Donovan, M. S., 153
- Doorey, N., 1158, 1159
- Dorans, N. J., 75, 164, 387, 408, 409, 410, 411,
412, 415, 419, 438n8, 600, 601, 610, 736,
743, 748, 750, 751, 752, 753, 756, 759,
760, 762, 763, 764, 765, 766–67, 768, 769,
770, 771, 773, 774, 775–76, 777, 778, 779,
780–81, 782, 783, 784, 785, 786, 788, 789,
792, 793, 796–97, 800, 801, 802, 803,
1205, 1547
- Dorsey, D. W., 131
- Dotter, D., 1135
- double administration procedure, 338
- Douglas, J. A., 669
- Downey, R., 536
- Downing, S. M., 130, 1314
- drag-and-drop type item, 528, 530f
- Drake, J. D., 353, 354
- Draney, K., 1068, 1069
- Drasgow, F., 403, 601, 618, 749, 1356, 1365,
1377, 1380–81
- Drew, J., 1281
- D studies. *See* decision studies
- Du, X., 928
- DuBois, P. H., 35, 37, 83, 518, 1270
- Duckworth, A. L., 425, 1345, 1347
- Dudley, N. M., 1342, 1352
- Dudley, R., 1203
- Dueber, D. M., 1376
- DuFour, R., 1039, 1040
- Dumais, S. T., 544
- Dunleavy, D. M., 1229
- Duolingo, 623–24, 636
- Duong, E. V., 547
- Duque, M., 606
- Durán, G., 423
- Duran, L. K., 976
- Durán, R., 120, 146, 148, 149
- Durlak, J. A., 1346, 1350
- Durrant, G., 1480
- Duschl, R. A., 1027, 1031, 1049, 1064, 1068
- Duster, A. M., 497
- Dweck, C. S., 1347, 1354–55
- Dwivedi, R., 622
- dynamic evaluation, 216–17
- Dynamic Language Learning Progression
project, 1067
- Dynamic Learning Maps, 1067
- EAP. *See* expected a posteriori estimator
- Early Childhood Longitudinal Study, 659
- Easton, J., 1137
- Ebbs, D., 122, 1436–37
- Ebel, R. L., 200, 298, 425, 850
- Ebner-Priemer, U., 1389
- Eccles, J. S., 127
- ECD. *See* evidence-centered design
- economic development, ILSAs and, 1434–35
- economic preference parameters, 1353
- The Economist* (magazine), 1209
- Edgeworth, Francis, 40, 67–68, 112n20, 1524
- Edley, C., Jr., 10
- EdTech Strategies, 587
- Education Act (2012), 1486
- educational attainment: cognitive skills rela-
tion to, 1434, 1469, 1471; globalization
relation to, 136; intelligence testing and,
112n28
- Educational Consolidation and Improvement
Act, 62
- educational disparities, 990; achievement
gaps and, 1117; assessment development
relation to, 980; between racialized
groups, 976–77; racism and, 977, 979;
segregation relation to, 978
- educational expectations, on-track labels and,
207
- educational games. *See* games and simulations
- educational opportunities, 197, 227–28
- Educational Opportunity Survey, 1119
- educational policy, 116, 1434, 1439; for digital
assessments, 1449; ILSAs relation to,
1425–26, 1467–70, 1473, 1484–88, 1490–
91; Improving America’s Schools Act, 62,
142, 442, 483, 977, 1122; international
assessments and, 140; stakeholders for,
1433, 1435, 1436. *See also* Elementary and
Secondary Education Act; Every Student
Succeeds Act; No Child Left Behind Act
- educational privatization, 1032
- Educational Psychology* (Thorndike), 73
- educational testing industry, 500–501, 516,
879, 1524, 1566
- Educational Testing Service (ETS), 80, 82,
89, 112n25, 403, 405, 412, 521, 534, 535,
540–42, 551, 585, 606–7, 620, 789, 983,
996, 1202, 1215, 1216, 1220, 1221, 1239,
1363, 1433, 1475
- Education Commission of the States, 89,
1128, 1133, 1211
- Education Longitudinal Study, 393
- education reform movements, 133, 134, 139
- Education Week* (newspaper), 8
- Edwards, A., 1196–97

- Edwards, B., 1445
 Edwards, M. C., 17–18, 665, 673, 686, 694, 1296
 Edwards, N. R., 1148
Edwards v. Aguillard, 7
 effectiveness, of score reporting, 901–2
 effective test length, 343
 Egan, K. L., 458, 460, 463, 859
 Eguchi, M., 619
 Ehrhart, M. G., 1385
 Eichinger, R. W., 1355, 1356
 Eichorn, L., 972
 Eid, M., 1360
80 Days (game), 532, 533f
 Eignor, D. R., 585, 794, 795
 Einhorn, H. J., 416, 421
 Ejima, C., 533
 Elby, A., 1068, 1069
 Elchert, D. M., 1267
 electronic health systems, 1320
 Elementary and Secondary Education Act (ESEA), 62–63, 133, 964; accountability in, 142; educational disparities and, 977; ELs in, 145; flexibility waiver and, 1141; NAEP relation to, 1131; Title I accountability requirements in, 1119
 Elementary and Secondary School Emergency Relief Fund, 6
Elements (Euclid), 1517
Elements of Generalizability Theory (Brennan), 69–70
Elements of Psychophysics (Fechner), 46, 47, 71
 Elias, S., 9
 Elledge, A., 1164
 Elliott, J. L., 965
 Elliott, S. N., 832, 1156
 Ellis, D. D., 408–9
 Ellis, J. L., 362
 Ellison, H. B., 1382
 Elman, B. A., 132
 Elmore, R., 1117, 1163
 Elo, A., 624
 Elo rating system, 624
 ELs. *See* English learners
 Elvira, M., 1375–76
 Elwert, F., 979
 EM. *See* expectation-maximization algorithm
 embedded standard setting, 722, 877–78
 Embretson, S. E., 203, 212, 445, 450, 500, 712, 742, 744, 746, 1292, 1305–6, 1421n4, 1551–52
 emission probability, 627
 Emmerich, W., 987
 emotional intelligence, 1345, 1353
 empirical ability distribution, variance and, 353
 empirical recovery, 1063
 employer surveys, 1191, 1239, 1242, 1343–44
 employment tests, 394, 1317; bias in, 403; certification and, 1273–74; cognitive ability and, 962; prediction invariance and, 419
 end-of course exams (EOC), 395, 1138–40
 engagement: assessment design and, 255; educator, 1168; motivation and, 982, 984; panelist, 873, 876; student, 11, 158, 494, 498, 1045, 1074, 1127, 1147, 1544; UDL and, 245; virtual agents and, 259
 Engelhard, G., 73, 240, 520, 523, 525, 740, 870, 874, 1045, 1532, 1533
 Engeström, Y., 153
 Englert, D., 1169
 Englert, K., 119, 120, 992
 English, D., 1061, 1128, 1129
 English language proficiency tests, 145, 146
 English learners (ELs), 118–20, 241; accessibility for, 967–68, 974–75; assessment development relation to, 976; construct-irrelevant variance and, 166; DIF and, 410; in ESSA, 145, 146–47; in K–12 education, 973–74; learning progressions for, 1067; NCLB and, 1164–65; SWDs as, 980–81; test accommodations for, 140, 143, 167, 517, 610, 974, 976; written testing and, 121–22
 English-speaking tests, 137
 entry-level practitioners, KSAs for, 224
 Environmental Protection Agency, 824
 EOC. *See* end-of course exams
 epistemologies, 116
 Equal Employment Opportunity Commission, 1280, 1282
 Equal Employment Opportunity Program Terminology, 437n6, 1269
 equal performance principle, bias and, 392
 equating, 91, 761t; Army Alpha and, 764; for Cooperative Achievement Tests, 765; data collection and, 738t, 770, 773; equipercentile, 58, 775, 780; IRT and, 765, 772, 779–80; linking relation to, 766–67, 769; measurement error and, 797; NEAT design and, 776–77, 778; raw-to-scale score conversions and, 785–86; reliability relation to, 768, 796; scale-aligning methods and, 770, 774–75, 783–84; scaling relation to, 16; score reporting and, 736, 760, 804; SEs of, 781–82; subpopulation invariance evaluations and, 782–83; true score and, 765, 767, 778–79, 780
 equating drift, 786
 equipercentile equating, 58, 774–75, 780
 equipercentile scale-aligning functions, 779
 equitability, of admissions testing, 16
 equity, 21, 125, 384; collaboration for, 245; in credentialing, 260–61; culturally responsive assessment relation to, 235–36, 237, 239; error variance and, 768; fairness and, 161, 392; learning progressions relation to, 156; linking and, 796–97; motivational, 1046; in score interpretation, 12; socio-cultural learning theories for, 252; summative assessments for, 1051; in test-based accountability, 1166; test scores relation to, 1167; test-taker characteristics relation to, 241; validity and, 235
 equity frame, for OTL, 8–9
 equity indicators, 10, 795, 1118, 1167
 equivalent evidence, 124, 154
 erasure detection methods, 519
 e-rater software, 535–36, 537f, 541
 Ercikan, K., 13–14, 116, 119, 121, 122, 123–24, 127, 131, 135, 136–37, 140, 149, 154, 156, 158, 161–62, 163, 164–65, 212–13, 235, 240, 241, 243, 254, 413, 455, 459, 551, 611, 613–14, 660, 724, 749, 928, 986, 988, 999, 1018, 1168, 1192, 1274, 1281, 1481
 Eren, O., 1149–50
 Erickson, K. A., 970
 Ericsson, K. A., 131
 error control, 1445
 error of approximation, 677
 error variance, 312–13, 318, 333, 348, 355; accessibility relation to, 965; ANOVA and, 313, 319; in Bayesian framework, 329; for composite scores, 296, 306; CSEM and, 326; CTT and, 295; equating and, 767; equity and, 768; IRT and, 316; linear composite and, 321; reliability and, 363; scaling and, 743; UG and, 303–4; variance-ratio definition and, 288–89
 Ertl, H., 1468
 ERW. *See* evidence-based reading and writing
 ESEA. *See* Elementary and Secondary Education Act
 Espin, C. A., 1053, 1055, 1057, 1059
 Espin, C. L., 1059
 ESSA. *See* Every Student Succeeds Act
 essay examinations, reliability for, 68–69
The Essentials of Mental Measurement (Brown), 45
 Estabrook, R., 1365, 1374
 Estimated Number and Percentage of U.S. High School Graduates Taking College Admissions Tests in 2017, 1212f
 Estimated Percentage of U.S. Postsecondary Institutions Using Selected Math and Reading Tests for Placement in Fall 2011, 1235t
 estimation, 279–80
 Estimators of Conditional Standard Errors of Measurement Based on a Single Test Administration, 336t
 ETS. *See* Educational Testing Service
ETS Guidelines for Developing Fair Tests and Communications, 395
ETS v. Hildebrant (2007), 1286
 Euclid, 1517
 eugenics movement, 13, 38, 45, 53, 59, 111n15; intelligence testing in, 35, 55–56, 92–93; Pearson and, 41–42, 110n3; racism in, 133; SAT relation to, 385
 Eugenics Record Office, 56
 Evanini, K., 540

- Every Student Succeeds Act (ESSA), 10, 32n7, 62, 143–44, 1050; accountability and, 1118, 1127–30, 1211, 1340; accountability testing and, 1114; alignment and, 1159; content standards and, 970, 1155–56; ELs in, 145, 146–47; growth tracking and, 1152; Innovative Assessment Demonstration Authority, 632; NAEP relation to, 1132; standardized tests and, 386; SWDs and, 1165–66
- evidence accumulation, 619
- evidence-based reading and writing (ERW), 1203
- evidence-centered design (ECD), 244, 445, 450, 909–10, 1551; design science and, 451; evidence model in, 482; KSAs and, 257–58; PAD and, 448, 452, 500; in TBA, 1305; TE and, 985; validity and, 551, 1292
- evidence model, in ECD, 482
- evidentiary argument: in PAD, 482–83, 485–86, 487, 493; for score interpretation, 485; TOA and, 484
- evidentiary reasoning, 450–51
- Ewers, N., 135
- Ewing, M., 455, 459
- The Examination of School Children* (Pyle), 53
- Example Item in English and Spanish to Illustrate Accessibility Features, 496f
- Example Items Used in Analyses, 700f
- Example Range Performance-Level Descriptors, 464t
- Examples of Classroom Formative Assessment, 1024t
- expectation-maximization algorithm (EM), 686, 688–89, 690
- expected a posteriori estimator (EAP), 309–10, 316, 317, 329, 1379, 1380; ability estimates and, 746, 747–48; IRT and, 675, 745, 746; proficiency scoring and, 322; vertical scaling and, 791
- expected error variance, 282–83, 311
- expected proficiency variance, 311
- expertise, test scores and, 550–51
- expert panels, 852–53, 854
- explainable AI, 622
- explanatory reporting, 912
- extended-interaction items, 593, 593f, 594, 598, 634
- extended multivariate GT, 306–7, 309
- external anchors, 776–77, 785
- extrapolation, 1302; in collaborative problem solving assessments, 260; for credentialing, 226–27; IUA and, 1303; test scores and, 249
- extrapolation inferences, 211, 214
- eye tracking, 254, 936
- Ezeife, A. N., 992–93
- F-1 visa program, 26
- facets, in GT, 299, 1527
- facial recognition software, racial bias in, 259
- facilitator, 19, 848, 856
- factor analysis, 349, 543, 628; covariance/correlation structure and, 676; higher order, 669; independence assumption for, 670–72; latent structure analysis relation to, 76; MIRT compared to, 666; multivariate normality and, 673
- factorial invariance, 414–15
- factorial validity, 199
- Fadel, C., 1348
- Faggen, J., 403
- fairness, 37, 384; accessibility and, 960, 962, 964–67, 999; of achievement tests, 397–98; in admissions testing, 22, 387, 404–5, 425, 426, 1242; assessment design relation to, 236, 961, 998–99; assessment development and, 598, 961, 982, 983; assessment performance and, 134–35; AVs and, 1372–73, 1374; BARS and, 1375–76; bias and, 163, 423, 436n2, 961, 1308; CLR and, 497–98, 966, 989; coaching relation to, 396; collaboration for, 245; construct-irrelevant variance and, 156, 218, 966, 997; construct underrepresentation and, 394; COVID-19 pandemic and, 795; in credentialing, 260–61, 999, 1269; cultural validity and, 992; decision-making and, 963; diagnostic testing and, 146; DIF and, 119, 122–23, 423, 985–86; in digital environments, 137; diversity and, 14–15, 135; equity and, 161, 392; forced-choice methods and, 1380–81; IUA and, 216; meritocracy and, 133, 138; PAD and, 470; performance assessment and, 1388; prediction invariance and, 419; in promotion exams, 1150; of remotely administered exams, 400–401, 402; respectful treatment and, 400; of score interpretation, 195, 196, 387; in selection decisions, 420–22; SJTs and, 1386; sociocultural context and, 120, 125, 136, 161–62, 239, 988; special needs and, 146–48; standardization and, 124, 135, 161, 164–65, 1303; of standardized tests, 384, 392, 1230; SWDs and, 123, 388; TBA and, 260, 589, 602–3, 634; TE for, 984; test accommodations and, 20, 145, 399, 424, 970, 997, 1363; test content and, 387, 393–94, 399, 409, 963, 964, 989; test scores and, 19, 392; test security for, 518; test-taker characteristics relation to, 241; test use and, 195–97, 218, 236, 400, 424–25; in UDA, 973; validation and, 389; validity and, 16, 34–35, 168, 193, 194–97, 229, 275n1, 387–89, 393, 408–9, 424, 552–53, 608, 724, 960, 1351
- FairTest, 1192, 1194
- faking, 1362, 1368, 1370–71
- Falk, A., 1353, 1388
- Falk, C. F., 1365, 1421n3
- Falmagne, J. C., 152, 623
- false negatives, 226, 340, 341; marginal category probabilities for, 342; true score and, 1317
- false positives, 226; cut score and, 1317–18; marginal category probabilities for, 342; marginal error rate for, 339–40, 341; in TOEFL, 234
- Faltis, 1071
- familiarity requirement, 390–91
- Family Educational Rights and Privacy Act (FERPA), 142
- family functioning, culture relation to, 126
- family-generated clone items, 749
- Family Institute, 139
- Fanelli, D., 656
- Faria, A., 1035
- Faridee, A. Z. M., 552
- Farmer, M., 968
- Farrell, C. C., 1034, 1036, 1041, 1050, 1136
- Farrington, C. A., 1352
- Farruggia, S. P., 1353
- Faulkner-Bond, M. M., 120, 121, 147, 976
- Fawcett, E., 6–7
- Faxon-Mills, S., 1047, 1050
- Feagin, J. R., 11
- feature development, 536
- feature evaluation, 618–19
- feature extraction, 542, 546, 547
- Fechner, G. T., 34, 35–36, 37, 46–47, 70, 71–72, 93, 112n22, 839, 1518–19, 1522, 1530
- Federal Aviation Administration, 1275
- federal government, 25–26
- federal legislation, 58, 63, 976–77. *See also specific legislation*
- Federal Trade Commission, 1280–81
- feedback: in assessment development, 1338n4; for classroom formative assessments, 1023, 1025, 1031–32; criterion-referenced, 861, 1036; diagnostic, 535; formative, 615, 626, 1029, 1297; grading as, 1042, 1044, 1045–46; learning progressions and, 1070; motivation relation to, 1026; normative, 850–51; proactive reciprocity for, 1027; for professional development, 1322; from SLOs, 1061–62; subscores as, 1295–96; teacher evaluation and, 1141–42
- Feigenbaum, M. D., 780
- Fein, P., 1234
- Feinberg, R. A., 353, 1296, 1308
- Feldon, D. F., 1228
- Feldt, L. S., 278, 279, 285, 291, 292–93, 295, 297, 318, 319, 325, 330–31, 333, 335, 360
- Feldt-Gilmer coefficient, 295
- Feldt's coefficient, 295, 297
- Fels study, 111n12
- Felton, E., 1203
- Feng, S., 1369
- Feng, Y., 632
- Feniger, Y., 139
- Ferdous, A. A., 849, 865, 872

- Féré, Charles, 47
 Fergus, E., 237
 Ferguson, A., 740, 1027, 1521
 Ferguson, C. J., 1346
 Ferguson, G. A., 1518
 Ferguson, P., 1046
 Ferguson Committee, 1521
 FERPA. *See* Family Educational Rights and Privacy Act
 Ferrando, P. J., 317
 Ferrando, S., 477–78
 Ferrara, S., 18, 131, 224, 230, 442, 456, 460, 463, 467, 469, 518, 555, 756, 829, 849, 851, 861, 877, 907, 908, 1032, 1199, 1310, 1312, 1319, 1547
 Festinger, L., 831, 863
 Feuer, M. J., 801–2, 1425
 Few, S., 924–25
 Fielding, A., 681
 Fields, R., 1234, 1236
 field testing, 473, 932–36
 Fife, D. A., 364
 Figlio, D. N., 1135, 1143, 1144, 1145
 Figueroa, 146
 Finch, H., 1309
 Fine, S. A., 1290
 Finkel, J. R., 558
 Finn, B., 1211
 Fischer, G. H., 1532
 Fischhoff, B., 1387
 Fischman, G. E., 898, 1467, 1469, 1484
 Fishbein, B., 1437
 Fisher, D., 1040
 Fisher, I., 56
 Fisher, R. A., 47, 68, 69, 77, 93, 687, 1530
 Fisher information function, 674, 675
 Fisher's identity, 687–88
Fisher v. University of Texas et al. (2016), 1201
 Fiske, D. W., 202, 1357
 fitted item factor model, 719
 fitting propensity (FP), 698–99
 Fitzpatrick, A. R., 81, 89, 278, 656, 747, 831, 875
Five Decades of Public Controversy Over Mental Testing (Cronbach), 384
 “five studies,” 35, 83
 fixed-form assessment, 474
 flagging policies, for test accommodations, 405–6
 Flake, Jeff, 1346
 Flanagan, A., 1122
 Flanagan, J. C., 58–59, 60, 91, 736, 742, 749, 750, 758, 760, 762, 765, 766, 769, 773, 782, 789, 802
 Flanagan-Guttman-Rulon coefficient, 292
 Fleiss' Kappa, 357
 flexibility waiver, 1141–42
 flexMIRT, 685, 703, 711–12
 Flier, L. A., 1319, 1322
 Flower, L., 553
 Floyd, G., 8, 32n4
 focus groups, 865, 931–32, 935–36, 1046
 Foegen, A., 1055
 Fogg, N., 1470
 Folley, S., 924–25
 Foltz, P. W., 358, 542, 544, 547
 forced-choice methods, 23, 1376, 1377, 1379–82, 1391
 foreign-language scoring engines, 558–59, 559t
 Forer, B., 217
 Forgash, E., 10
 formal documents, 129
 Forman, E. A., 1028
 Formative Assessment Design Cycle, 1040
 formative assessments, 1021t; assessment for teaching and learning relation to, 1020–21; BEAR Assessment System and, 1069; CBM and, 1058, 1059; classroom questions relation to, 1027–28; common, 20, 1038–39, 1040–41; continuous testing and, 1307; instructional decisions relation to, 1050; interim assessment compared to, 1076, 1079–80; learning progressions and, 1063, 1066–67, 1068, 1069–70, 1071; MATHia and, 1073; norms relation to, 1030; professional development and, 1029–30, 1039; rubrics in, 1031; score reporting for, 917, 925, 926, 942; summative assessments compared to, 1019, 1076; SWDs and, 1022. *See also* classroom formative assessments
 formative feedback, 615, 626, 1029, 1297
 form parallelism, 292
 Forsyth, R. A., 759
 Forte, E., 1158–59
 Forte Fast, E., 916, 935
The Forum (magazine), 85
 Foster, D., 1305, 1307–8
 4Sight benchmark tests, 1037
 four-parameter IRT, 112n23
 Fourteenth Amendment, 149, 150, 1280
 Fowler, F. J., 1358
 Fowles, M. E., 1228
 Fox, E., 1533, 1542, 1543, 1547
 FP. *See* fitting propensity
 FR. *See* free-response items
 fractional scores, 357
 A Framework for K–12 Test-Based Accountability Purposes and Mechanisms, 1116f
 Framework Relating Formative Assessment Strategies to the Formative Assessment Process and Its Agents, 1021t
 Framework to Guide Workshop Design and Gather Validation Evidence to Support Claims about the Interpretation of Cut Scores, 857t
 framing effects, 843–44, 867
 France, 1486
 Frank, B., 360
 Frankenber, G., 978
 Franks, V., 1195, 1196
 Frary, R. B., 1045
 Frederiksen, J. R., 460
 Frederiksen, N., 590, 1340
 freedom of speech, 10
 Freeman, M. F., 753
 free-response items (FR), 306, 307, 309
 Fremer, J. J., 518
 French, B. F., 1309
 French, J. W., 1387
 Freund, P. A., 1362, 1385
 Frey, C. B., 1345
 Frey, N., 1040
 Friesen, J. B., 992–93
 Frisbie, D. A., 1042
 Frisch, Ragnar, 77, 79
 Froese-Germain, B., 139–40
 frog dissection simulation, 530, 531f
 Frohbieter, G., 1033
 Fu, J., 632, 1377
 Fuchs, D., 1053, 1055, 1057
 Fuchs, L. S., 969, 1053, 1054, 1055, 1056, 1059
 Fuchsman, D., 1147
 Fuess, C. M., 87, 1202
 funding, 25, 443–44
 Furtak, E. M., 118, 1020, 1028, 1029, 1030, 1039, 1040, 1064, 1068, 1069–70
 Furuta, J., 1195
 Futagi, Y., 536
 Futia, G., 622
 G20, 622
 Gaddis, S. M., 1144–45
 Gaertner, M. N., 423, 878
 Gaff, J., 1199
 Galla, B. M., 1043
 Gallagher, A., 597–98, 603, 984
 Galli, L., 520
 Gallimore, R., 1040
 Galton, F., 36, 38–40, 41–42, 47, 51, 55, 64, 67, 71, 72, 73, 92, 93, 111n13, 852, 893n6, 1519–20, 1522, 1540
 Galton Biometric Laboratory, 64
 games and simulations, 23, 257, 259, 494, 529, 530; AI in, 624, 637; cheating in, 520; cognitive domain model in, 635–36; for credentialing, 1294; data collection and, 660; 80 Days, 532, 533f; frog dissection, 530, 531f; ILSAs with, 1482; intelligent tutoring systems compared to, 625; IUA for, 257–58; reliability and, 366–67; serious-based games, 530, 532; in TBA, 254
 gaming behavior, 1147, 1151
 Gamlem, S. M., 1027
 Gan, N., 680
 Gándara, P., 148
 Gao, J., 1213, 1298
 Gao, L., 1388
 Gao, R., 766
 Gao, X., 1074
 García, G. E., 146

- Garcia, G. N., 974
 Garcia-Molina, H., 521
 Gardner, J., 995
 Garrett Holliday, B., 977
 Gaussian kernel smoothing, 775
 Gay, G., 446, 497
 gay liberation movement, 10
 Gaza, 10
 Gebril, A., 117
 GED. *See* General Educational Development certificate
 Gee, J. P., 119, 632
 Geiser, S., 407
 Geisinger, K. F., 962, 1316, 1317–18, 1319
 Geist, L. A., 970
 Geldhof, G. J., 350, 355
 Gelman, A., 697
 Gelman, S. A., 116
 gender, 987, 1471
 General Accounting Office, 130
 General Educational Development certificate (GED), 1263n8, 1341
 general intelligence, 45, 56, 1520
 “General Intelligence’ Objectively Determined and Measured” (Spearman), 43
 generalizability coefficient, 288, 303
 generalizability studies (G studies), 298; multivariate GT and, 304–5; variance in, 299, 300–301, 302, 307, 309
Generalizability Theory (Brennan), 298
 generalizability theory (GT), 15, 36, 68, 112n19, 1526; absolute error variance in, 327, 820n1; AVs and, 1372; classification accuracy and, 348; coefficient alpha and, 381n1; CTT and, 63, 111n18, 297, 298, 1527, 1528; error variance and, 355; facets in, 299, 1527; IRT and, 357, 1529, 1530; multivariate, 304–7, 309, 333–34; Personal Potential Index and, 1367–68; person-by-task interaction in, 163; random-item models and, 681–82; rater reliability and, 356, 365; reliability coefficients and, 364; reliability in, 66–67, 92, 214, 278, 280–81, 297–98, 361; replication in, 300; SJTs and, 1385–86; speeded assessments and, 354; test scores and, 68, 213–14; true score in, 282; univariate, 69, 298, 328–29; universe score in, 67, 302, 1531; validity and, 197–98; variance in, 360
 generalization inferences, 211, 1302, 1303
 generalized graded unfolding model (GGUM), 1378
 generalized modeling framework, 685
 generalized partial credit model (GPCM), 81, 667, 682, 719, 1360–61
 generalized Spearman-Brown formula (GSB), 293
 General Language Understanding Evaluation, 549
 general-purpose reliability coefficient, 296
 general test factor, 352
 Generating Performance-Level Descriptors for the Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores Partnership, 481f
 Gensowski, M., 1342
 Georgia Department of Education, 632
 Gerasimova, D., 216
 Gergen, K. J., 656
 Gerst, M. D., 946
 Gersten, R., 1055
 Gesel, S. A., 1057
 Gewertz, C., 406, 1139
 GGUM. *See* generalized graded unfolding model
 Ghai, W., 526
 Ghandour, G., 360
 Gialluca, K. A., 1310, 1311
 Gianinazzi, M. E., 357
 Gibbons, R. D., 320, 666, 676
 Gibson, S., 1145
 Gierl, M. J., 131, 147, 457, 542, 583, 943, 1292, 1305, 1551–52
 gifted children, 54
 Gigerenzer, G., 845, 848
 Gilford, D. M., 1432
 Gillham, N. W., 72
 Gilmer, J. S., 295
 Gilmour, A. F., 118, 1142–43
 Ginsburg, A. L., 1121
 Gippes, C., 1025
 Giraud, G., 869
 Gitomer, D. H., 623, 631, 983, 1018, 1027, 1031, 1049, 1073
 Glas, C., 749, 798, 1457
 Glaser, R., 131, 338, 449, 1118–19
 Glaser, Robert, 1131
 Glass, G. V., 1318
 Glazer, N., 521
 Glazerman, S., 1135, 1154
 Gleser, G. C., 69
 Gleser, L. J., 1377
 globalization, 136, 140, 1488
 Glynn, J., 423
 GMAC. *See* Graduate Management Admission Council
 GMAT. *See* Graduate Management Admission Test
 GMAT Focus Edition, 1220
 GMPP. *See* Growth Model Pilot Program
 Gneezy, U., 1240
 Gobert, J. D., 624, 1074
 Gochyyev, P., 1070
 Godambe, V. P., 77
 Goddard, Henry, 36, 51–52, 56, 93, 110n7, 110n9, 1523
 Godfrey-Smith, P., 1577n2
 Goertz, M. E., 250, 1035–36
 Goffin, R. D., 1375
 Goggin, W. J., 423
 Golan, 148
 Goldberg, J., 1209
 Goldberg, L. R., 1342, 1352, 1356, 1357
Golden Rule Insurance Company v. Washburn (1984), 403–4
 Golden Rule procedure, 403–4
 Goldhaber, D., 1143, 1155
 Goldhammer, F., 1480
 Goldin, I., 1067
 Goldman, S. R., 127
 Goldrick-Rab, S., 127
 Goldstein, A. A., 845, 969
 Goldstein, D., 8, 26, 454
 Goldstein, H., 681, 1472
 Goleman, D., 1345
 Golub, G. H., 543
 Golub-Smith, M. L., 754
 Gonchar, M., 995–96
 Gong, B., 124, 194, 242, 249, 1051
 Gonzalez, J., 765–66, 773
 Goodman, D. P., 443, 919, 927, 931, 937, 938, 943
 Goodman, M. J., 1471
 goodness-of-fit tests, 676, 678, 698
 Goral, D. P., 1067
 Gordon, E. W., 261
 Gordon Commission, 261
 Gorin, J. S., 137, 1292, 1546, 1551–52
 Gosling, S. D., 1390
 Gotch, C. M., 937
 Gottfredson, Linda, 57
 Gottlieb, M., 146
 Gould, S. J., 52, 57, 132, 133, 235, 385, 1564
 Government Accountability Office, U.S. (USGAO), 971–72
 Gowan, P., 485, 1030, 1070
 GPA. *See* grade point average
 GPCM. *See* generalized partial credit model
 Grade 4 Mathematics Policy-Level Definitions, 461f
 graded response model, 80, 682–83
 grade inflation, 1195, 1198
 grade point average (GPA), 214, 1192–93; academic performance relation to, 1224, 1232; ACT relation to, 1226; admissions testing and, 417; Big Five personality model relation to, 1368–69; correlation coefficient and, 1230; forced-choice methods and, 1381; GMAT relation to, 1227–28; GRE relation to, 1219; LSAT and, 1220; MCAT and, 1222; prediction invariance and, 418; SES and, 393; test scores and, 1263n3; TOEFL relation to, 233–34
 grade promotion exam, 1140–41
 grading, 20, 1044; assessment literacy and, 1045; motivation relation to, 1046; validity of, 1041, 1042–43
 Graduate and Professional Admissions Tests, 1216f
 Graduate Management Admission Council (GMAC), 1213, 1215, 1220, 1227, 1230

- Graduate Management Admission Test (GMAT), 405, 585, 1215; academic performance relation to, 1227–28, 1231–32; CAT for, 1213; COVID-19 pandemic and, 1191, 1220; DIF and, 987; forced-choice methods and, 1381
- Graduate Record Examination (GRE), 17, 585, 996, 1219; academic performance relation to, 1228–29, 1232; adaptive testing and, 50, 610, 1218; CAT and, 585, 1213; forced-choice methods and, 1381; scale scores for, 750, 754, 1219; test accommodations for, 968
- graduate school admissions, 1213, 1214–15
- graduation rates, 1129, 1149
- graduation tests, 388, 1116–17
- Graesser, A. C., 1555
- Graf, E. A., 1063
- Graham, J. M., 349, 528
- Graham, M., 524
- Graham, P. A., 85
- Graham, S., 561, 1025, 1031, 1355
- The Graide Network, 402
- grammar schools, written testing in, 84
- Graphical Plate Representation of Latent Dirichlet Allocation, 545f
- Gratz v. Bollinger (2003), 1200
- Gray, B. M., 1304
- GRE. *See* Graduate Record Examination
- GRE Analytic Writing Assessment, 256–57, 618
- Great Society initiative, 62, 88
- Green, A., 538
- Green, B. F., 76, 91, 126, 310, 352, 393, 1361
- Green, D. R., 850
- Green, S. B., 349
- Greene, Jennifer, 448
- Greene, J. P., 1141, 1149
- Greenfield, P. M., 116
- Greeno, J. G., 126–27
- Greenwald, A. G., 1388
- Gregg, N., 517, 969, 970
- Gregory, A., 237
- Greiff, S., 614
- Grek, S., 140, 1484
- Gresalfi, M. S., 126–27
- Griesemer, J. R., 122
- Griffin, P., 927, 942
- Griffith, R. L., 1239
- Grissmer, D., 1122
- Grissom, J. A., 1147
- grit, 1347
- Grodner, A., 1390
- Grodsky, E., 1148, 1149
- Gross, A. L., 422
- group-based variance, 163
- group-by-task interaction, 162, 163
- group discussions, 871, 875–76
- group interaction submodel, 829, 831–32, 856
- group-level norms, 758
- group median cut score, 893n6
- group reports, 920; circles of interest for, 921, 922–23; credentialing and, 901; for K–12 assessment, 900
- group-score assessments, 601, 1425
- Grove, A., 499
- Grover, R., 135, 164
- growth mindset, 1026, 1347–48, 1351, 1359
- growth model analyses, 230
- Growth Model Pilot Program (GMPP), 1114, 1151, 1152, 1153
- growth norms, 758, 793
- growth tracking, 1151–53
- Grumm, M., 1388
- Grund, S., 1509n33
- Grünwald, P. D., 656, 698
- Grutter v. Bollinger (2003), 1200, 1214
- GSB. *See* generalized Spearman-Brown formula
- G studies. *See* generalizability studies
- GT. *See* generalizability theory
- Guenole, N., 1384
- Guidelines for Technology-Based Assessments, 550
- Guilford, J. P., 199, 1518
- Gulino v. Board of Education (2002), 1273, 1281–83
- Gulliksen, H., 37, 42, 43, 65, 69, 199, 200, 286, 353, 743, 765, 777–78, 784, 789, 1524
- Guo, F., 345–46
- Guo, H., 137, 595, 614, 615, 617, 786, 1384
- Guo, X., 524
- Gurkan, G., 1479
- Guskey, T. R., 1042
- Gustafsson, J. E., 1475
- Gutiérrez, K. D., 396
- Gutiérrez, R., 498
- Guttman, L. A., 294, 740, 1373
- Gwet, K. L., 359–60, 1367
- Haberman, S. J., 605, 759, 775, 785, 801, 803, 1362–63, 1383
- Hackett, R. K., 403
- Hackmann, D. G., 1128
- Haderlein, S. A. K., 1134, 1135
- Hadwin, A. F., 1022
- Haenszel, W., 411
- Haertel, E. H., 148, 208, 209, 210, 214, 217–18, 219, 220, 227, 230, 244, 278, 279, 285, 292, 351, 357, 360, 361, 367, 368, 445, 483, 743, 802, 1026, 1116, 1118, 1120, 1292, 1297, 1305, 1507n14, 1525, 1546
- Haertel, G., 909–10
- Hafner, J. C., 1031
- Hafner, P. M., 1031
- Haghighat, E., 423
- Hakuta, K., 143, 146–47, 973, 974, 978
- Haladyna, T. M., 583, 983, 994
- Haley, David, 77, 745
- Haley, S. M., 352
- Hall, E., 1061, 1375
- Hall, G. Stanley, 51
- halo effect, 847, 1368, 1370
- Hamaker, E. L., 350
- Hambleton, R. K., 19, 77, 79–80, 83, 110n8, 119, 147, 224, 226, 278, 338, 756, 828, 851, 852, 853, 854, 856, 859, 860, 861, 866, 868, 879, 896–98, 901, 902, 907, 908, 910–11, 912, 914–15, 916, 918–19, 923, 927, 928, 929–32, 934–37, 938, 943, 944, 945, 946–47, 1295, 1310, 1313, 1315, 1316–17, 1318–19, 1547, 1548
- The Hambleton and Zenisky Model for Score Report Development, 19, 929f, 929–32, 934–37
- Hamilton, A., 1077, 1114, 1118, 1120
- Hamilton, L. S., 21, 119, 120, 130, 898, 1146, 1160, 1161, 1165, 1166, 1340–41, 1345, 1358, 1366
- Hammond, Z. L., 497
- Hamner, 556
- Hamp-Lyons, L., 234
- Hampson, S. E., 1371
- Han, C., 1264n11
- Handbook of Modern Item Response Theory (Hambleton and van der Linden), 80
- Handbook of Test Development (Lane), 448, 470, 937
- Handcock, G. R., 349
- hands-on tasks, person-by-task interaction and, 163
- handwriting quality, 739, 756
- Han dynasty, 35
- Haney, W., 87, 149
- Hanford, E., 454
- Hansen, E. G., 165
- Hansen, M., 17, 694, 697
- Hansen, M. A., 866
- Hansen, M. H., 1445
- Hanson, B. A., 318, 340, 343, 765, 768–69, 778, 1298
- Hanushek, E. A., 1051, 1077, 1132–33, 1144–45, 1470, 1471
- Hao, J., 612, 1388
- Hao, S., 346, 347
- Harik, P., 256, 871, 1312
- Harlen, W., 1021
- Harmes, J. C., 528, 1293–94
- Harold, J., 946
- Harris, C. J., 456, 500, 1224, 1298, 1526
- Harris, D., 1145
- Harris, D. J., 15, 367, 743, 745, 749, 765, 798–99
- Harris, D. N., 1135, 1139, 1148
- Harris, E. A., 8
- Harris, L. R., 1046
- Harris, M. J., 1025
- Harris, M. M., 1357, 1367, 1369
- Harrison, D. A., 854, 876
- Harry, B., 147
- Hart, B., 45
- Hart, M., 423

- Hart, R., 443
 Hartocollis, A., 8–9, 27, 1195
 Hartz, S. M., 669
Harvard Educational Review (journal), 56
 Harvard University, 10, 26–27, 1200–1201
 Harvey, A. L., 515, 855
 Harwell, D., 259
 Haskins, C., 654n6
 Hastie, R., 678, 841
 Hastings, R., 501, 1135
 Hastings, W. K., 686
 Hastings Robbins–Monro algorithm (MH–RM), 686, 688, 711–12
 Hattie, J., 626, 903, 906, 908, 914, 931, 937, 943, 944, 945, 947, 1022, 1023, 1025, 1026, 1031, 1348, 1369
 Hatzenbuehler, M. L., 138
 Hau, K. T., 1306
 Hauenstein, N. M., 524
 Hauer, K. E., 1319, 1321
 Hauser, R. M., 148, 1150
 Hawaii State Department of Education, 1169
 Hawkins, D., 1198
 Hawkins, R. D., 1302
 Hayes, B., 839
 Hayes, J. R., 553
 He, J., 1365, 1373, 1478
 Headrick, T. C., 754
 Head Start, 62
 Heaney, K. J., 969
 Hebert, M., 561
 Hecht, 1316
 Heckman, J. J., 1341
 Hedeker, D. R., 320, 666
 Hedges, L. V., 125
 Hedlund, J., 1382, 1386
 Heffernan, C., 627
 Heffernan, N., 627
 Hegarty, M., 931, 945–46
 HEIghten suite, 1239
 Hein, S. F., 844, 860, 867, 868, 869, 873
 Heinrich, C. J., 1144
 Heinrich, J., 161
 Heit, E., 839
 Hejri, F., 974
 Hemelt, S. W., 1144, 1148, 1149
 Henchy, A. M., 355
 Henderson, S., 1037
 Hendrick, I. G., 146
 Hendrickson, A., 466, 467, 489
 Henly, G. A., 592
 Henly, S. J., 676
 Henson, R. A., 668
 Heppen, J., 1035
 Heredia, S. C., 1021, 1038, 1039, 1040, 1064, 1068, 1069–70
Hereditary Genius (Galton), 38, 52
 Heritage, J., 1022, 1074
 Heritage, M., 485, 1022, 1025, 1029, 1030, 1062, 1063, 1067, 1070, 1074
 Herman, J. L., 148, 168, 445, 1031, 1118, 1120, 1158
 Hernandez, D. J., 1140
 Hernández, R. A., 1127
 Hernández-Torrano, D., 1470
 Herold, B., 401
 Herrington, C., 1145
 Herrnstein, Richard, 57
 Hertz, N. R., 854, 1287, 1288, 1289–90
 Hess, K. K., 1064
 Heubert, J. P., 1150
 heuristics, 844, 846–47, 860, 875
 Hewlett Foundation, 556
 Hewlett trials, 534, 556–57
 Heyneman, S. P., 140, 1432, 1468, 1487
 Hezlett, S. A., 1232
 Hickman, L., 552–53
 Hicks, L. E., 1377
 A Hidden Markov Model for Estimating Student State Over Time, 627f
 hierarchical diagnostic classification models, 1075
 hierarchical latent trait estimation, 353
 hierarchical linear model (HLM), 678, 681, 702, 716, 718
 Hierarchical Linear Model Results Based on Test Sum Scores, 703t
 hierarchical models, 34, 694, 1356
 Hieronymous scaling, 791
 Higgins, C., 1366
 Higgins, D., 536, 541
 Higgins, K. M., 1031
 higher order factor analysis, 669
 high-minority schools, COVID-19 pandemic and, 3–4
 High School and Beyond survey, 418
 high school exit exams (HSEEs), 1138–40, 1148–49
 high school grade point average (HSGPA), 1192–93, 1224–25, 1227; ACT relation to, 1197–98, 1226; admissions testing and, 1199; test-optional policies relation to, 1196
 high school summative assessments, 1130
 high-stakes decisions, 140, 150
 high-stakes outcomes, 962
 high-stakes purposes, 1155
 high-stakes testing, 141, 521, 962, 1160; admissions testing as, 1222; cheating and, 518; cut score in, 1297; discrimination in, 149; self-report ratings relation to, 1366; test security for, 1305, 1307; VAMs and, 1155
 Hilgard, E. R., 52
 Hill, H. C., 1146
 Hill, P. W., 1068
 Hill, R., 1152, 1153
 Hilton, M. L., 1238, 1345
 Hirschman, A. O., 1134
 Hislop, H., 1486
 Hiss, W., 1195, 1196
 historical thinking skill, 459
 history classes, DIF and, 413
 Hitt, C., 1388
 HLM. *See* hierarchical linear model
 Ho, A., 21, 148, 217–18, 227, 386, 483, 586, 616, 657, 658, 659, 793, 802, 962, 977, 1116, 1126, 1143, 1151, 1152, 1154, 1155, 1161, 1190, 1211, 1390
 Hochbein, C., 1043
 Hofer, G., 1385
 Hoff, D. J., 1122
 Hoffer, T. B., 1153
 Hofferage, U., 848
 Hofstee, W. K. B., 870
 Hofstee method, 870, 1312
 Hogg, R. V., 436n2
 Holahan, C. K., 54
 Holcombe, R., 1160, 1162
 Hölder, O., 740
 holistic review, 1199–1200, 1214–15
 Holland, D. F., 355
 Holland, P. W., 164, 218, 404, 408, 411–13, 419, 600, 736, 760, 762, 763, 764, 765, 766–67, 768, 769, 770, 771, 772, 773–74, 775–76, 777, 779, 780, 781, 782, 785, 786, 789, 790, 792, 793, 796–97, 800, 802, 1151, 1308, 1534, 1547
 Hollenbeck, 1055
 Hollie, S., 497
 Holmboe, E. S., 1304
 Holme, J. J., 1135, 1148, 1149
 Holmes, O. W., 56
 Holmgren, D., 88
 Holzinger, K. J., 45
 Hommel, K. A., 127
 homogeneity assumption, bias and, 123
 Hong, J. C., 455, 493
 Hontangas, P. M., 1376, 1379, 1380
 Hood, S., 195, 234, 236
 Hooker, Joseph, 38, 39
 Hooper, L., 911
 Hooper, M., 1475
 Hoover, E., 407
 Hoover, H. D., 755, 791
 Horkay, N., 587
 Horn, J. L., 1377, 1378
 horn effects, 847
 Horvath, 1383
 Hoskens, M., 763, 767, 793
 Hosp, J. L., 1056
 Hosp, M., 1056
 hot spot item, 528, 529f
 Hough, H., 1137, 1362
 Houts, C. R., 685, 705, 706
 Howard, R. A., 617
 Howell, J., 1196–97
 Howell, W. G., 1137
 Howson, P., 538
 Hoxby, C., 1211
 Hoyt, Cyril, 111n17, 298
 HSEEs. *See* high school exit exams

- HSGPA. *See* high school grade point average
- Hsieh, H., 122
- Hsu, W.-N., 539
- Hu, J. C., 400
- Hu, L. T., 678
- Hua, C., 1533
- Huang, S., 675, 681, 720, 722
- Huang, W., 80
- Hubbler, S., 133
- Huberty, T. J., 127
- Hubler, S., 1322
- Hubley, A. M., 217, 254, 1481
- Huddleston, E. M., 82
- Huebeck, E., 134
- Huebert, J. P., 148
- Huff, K., 16, 209, 244, 442, 443, 445, 448, 454, 456–57, 458, 460, 461, 468, 475, 476, 477–78, 480, 489, 500, 515, 583, 832, 859, 877, 910, 961, 982, 992, 1222, 1291, 1293
- Huggins-Manley, A. C., 552–53
- Hughes, D. J., 1553
- Hughes, G. B., 116, 117
- Hughes, K. R., 525
- Hughes, T., 392
- Hulleman, C. S., 1348
- Hullman, J., 946
- human capital theory, 1341, 1345–46, 1425, 1433–34, 1506n2
- human judgment, for standard setting, 823, 828–29, 834
- human proctors, 402, 1307
- human raters, 255, 521, 551, 556; automated scoring compared to, 358–59, 538, 558, 562, 620; for CAT, 366; for ILSAs, 1482; process data and, 616; resolved score and, 557; transparency and, 554, 555; for validation, 621
- human resources, international assessments and, 138
- human-scored responses, for reliability, 528
- Humphreys, L. G., 416
- Hunt, M., 1135
- Hunter, J. E., 1389
- Hunter, S. B., 1146–47
- Hurford, D. P., 454
- Hurtz, G. M., 516, 854, 1311
- Hurwitz, M., 231, 1198
- Husek, 1298
- Husén, T., 135, 1431
- Hussein, 749
- Hutt, E., 1116, 1117, 1130, 1132, 1134
- Huynh, H., 340, 343, 345, 759, 1298
- hybrid classes, COVID-19 pandemic and, 5
- Hyman, J., 231–32, 1211
- Hypothetical Comparison of Intended Performance Levels and Item Scale Placement, 478f
- Iacus, S. M., 605
- IALS. *See* International Adult Literacy Survey
- IBM Institute for Business Value Global Country Survey, 1238
- iBT. *See* Internet-Based Test
- ICC. *See* item characteristic curve
- ICCS. *See* International Civic and Citizenship Education Study
- ICILS. *See* International Computer and Information Literacy Study
- IDB Analyzer. *See* International Database Analyzer
- IDEA. *See* Individuals With Disabilities Education Improvement Act
- ID Matching, 842–43, 844–45, 849
- IEA. *See* International Association for the Evaluation of Educational Achievement
- IEA–ETS Research Institute, 1465
- IEP. *See* individualized education program
- IES. *See* Institute of Education Sciences
- Illustrating Mechanisms for Inflation When Test Developers Sample Predictably From Content Domains, 1162f
- ILSAs. *See* international large-scale assessments
- Immekus, J. C., 1033
- immigration laws, 56
- impact data, 835, 850–51, 861
- Impara, J. C., 851
- imperial civil service exams, 15–16
- Improving America's Schools Act (1994), 62, 142, 442, 483, 977, 1122
- Inchley, J., 1358
- inclusion, 143, 144–45
- independence assumption, 670–72, 674
- Indigenous language, 975
- Indigenous populations, 120, 135; accountability relation to, 140; cultural validity for, 993
- individual differences research, 1519
- individualized education program (IEP), 972
- individual panelist cognitive processing sub-model, 829, 831, 833
- individual proficiency estimates (P method), 346–47
- Individuals With Disabilities Education Improvement Act (IDEA), 142, 964, 1164; educational disparities relation to, 980; inclusion in, 143; special needs and, 146, 147–48; state testing and, 219; test accommodations and, 969
- inferential evidence, 1066
- Inferential Links in the Assessment Use Argument, 232f
- information processing, 19, 897–98, 943
- information technology, 131, 136–37, 1071–72, 1076–77
- infrastructure, 582, 584, 607, 634
- Ing, M., 252, 1048
- inheritance, 72
- Innovative Assessment Demonstration Authority, ESSA, 632
- innovative assessment designs, 17, 254, 634
- innovative item types, 597, 599, 610, 1293–94
- Inoue, A. B., 996
- Inq-ITS, 624, 636
- Inside Higher Ed, 1240
- Institute for Credentialing Excellence, 1293, 1294, 1300, 1310–11, 1322
- Institute for Education Sciences, 555, 582, 1465
- Institute for Experimental Psychology, University of Leipzig, 47
- Institute of Education Sciences (IES), 26, 927, 1134, 1345
- instructional alignment, 1146–47
- instructional decisions, 1050, 1059, 1062
- instructional evidence, 1066
- instructional methods, learning outcomes and, 148
- instructional planning, 230, 844, 926, 1018, 1038–39
- instructional relevance, 979
- instructional sensitivity, 1156–57
- instructional strategies, 1057, 1060–62
- instructional validity, 1156–57
- instrumentation, 24, 1538, 1545–46, 1549
- Integrated Judgment procedure, 827, 837
- intellectual disabilities, 48–49; ELs relation to, 147; state testing and, 219–20
- intelligence quotient (IQ), 34, 52, 54, 111n13, 1343; cognitive skills and, 1345; fairness and, 387; racism and, 56–57; special education and, 388
- intelligence testing, 51, 87, 91, 111n14, 1520–21, 1522–23; bias in, 50, 56; Binet and, 36, 37, 49–50; Binet-Simon, 385, 739, 741, 759; cross-cultural, 155; educational attainment and, 112n28; in eugenics movement, 35, 55–56, 92–93; factorial invariance in, 414; Goddard and, 52; large-scale, 52–55; standardized tests for, 386
- Intelligent Essay Assessor, 547
- intelligent tutoring systems, 623–24, 626; games and simulations compared to, 625; real-time learning interactions and, 632
- interaction hypothesis, 609–10
- interim assessment, 20, 444, 498, 1078; assessment design for, 1033; CAT for, 1073; data collection and, 1034–36, 1037–38; decision-making relation to, 443; formative assessments compared to, 1076, 1079–80; NCLB and, 250, 1032, 1074; TOA for, 484, 484f, 485; validation relation to, 250, 251–52; validity of, 252, 1033–34
- internal anchors, 776–77
- Internal Medicine Certification exam, 1295–96
- internal structure, 211, 214, 909, 1065; DIF and, 240; validity and, 199, 239, 908
- International Adult Assessments, 1430f
- International Adult Literacy Survey (IALS), 1430

- international assessments, 131, 138, 140
- International Association for the Evaluation of Educational Achievement (IEA), 975, 1120, 1424, 1431–32, 1466; IDB Analyzer, 1464–65, 1470, 1509n42; ILSAs of, 1427; PUF and, 1462
- International Baccalaureate, 446
- International Civic and Citizenship Education Study (ICCS), 1431, 1507n9
- International Computer and Information Literacy Study (ICILS), 1431, 1507n8
- International Database Analyzer (IDB Analyzer), 1464–65, 1470, 1509n42
- International Health Exhibition, 39, 1519
- international large-scale assessments (ILSAs), 23–24; adaptive testing and, 1447–48; assessment design of, 1438; assessment development of, 1440–41; comparability and, 1436–37, 1441, 1448, 1472, 1477, 1481; as computer-based assessment, 1439; consequential validity of, 1484–85, 1491; data cleaning for, 1460; data collection and, 1424–25, 1439, 1449, 1466, 1471, 1474–75, 1488; data security for, 1465; as digital assessments, 1447, 1455, 1477–78, 1483, 1489–90; economic development and, 1434–35; educational policy relation to, 1425–26, 1467–70, 1473, 1484–88, 1490–91; human raters for, 1482; item calibration and, 1451; latent regression model and, 1459; linking and, 1453–55; for literacy, 1427, 1429–30, 1432, 1438; measurement invariance and, 1436–37, 1456–57, 1478; motivation relation to, 1480; NAEP and, 1424, 1432–33, 1489–90; PBA and, 1446, 1477; policy makers relation to, 1470, 1473, 1485–86, 1489; population sampling and, 1443–45; PUF for, 1461–62; test accommodations for, 1448–49, 1479; translation procedures for, 1442–43, 1472, 1479; validation of, 1437, 1449–50; validity of, 1476, 1481. *See also* large-scale testing; *specific assessments*
- International Organization for Standardization (ISO), 1465
- International Personality Item Pool, 1356
- international students, at Harvard University, 27–28
- International System of Quantities, 1518
- International Test Commission (ITC), 121, 124, 550, 906
- Internet-Based Test (iBT), 233
- interpersonal skills, 1350–51, 1357, 1366
- Interpretation and Use Argument and Validity Argument Components for Culturally Responsive Assessments, 247t, 248t
- interpretation model, 151, 157–58, 159, 159f, 167
- Interpretation of Educational Measurements* (Kelley), 65
- interpretation/use argument (IUA), 14–15, 192, 194, 218, 907; admissions testing and, 231; AES and, 256–57; bias relation to, 261; comparability and, 242; confirmation bias and, 209, 216; consequences and, 229; for culturally responsive assessment, 235, 237, 246, 247t, 248t; data logging and, 255; extrapolation and, 1303; for games and simulations, 257–58; for interim assessment, 250, 251; for large-scale testing, 239–40; licensure and, 1302; score interpretation and, 253; score reporting and, 930, 947; scoring inferences and, 249; in test interpretation, 208–12; in test scores, 197–98; as TOA, 219, 220–21; validity and, 193, 203–4, 206, 208, 212, 219, 222, 228, 238, 251, 253, 260, 928; validity argument and, 234
- interpretive guides, for score reporting, 931
- interrater reliability, 492
- intersectorality, DIF and, 986
- interval estimation, 286, 289, 337, 349
- interval measurement, 741–42
- interval properties, test performance with, 755
- interviews, for score reporting, 935–36
- intraclass correlations, 311–12, 317, 352
- intrapersonal skills, 1350–51, 1357, 1366
- Ion, A., 1363
- Iowa Academic Meet, 88
- Iowa Test of Basic Skills, 86, 789
- Iowa Tests of Educational Development (ITED), 750–51, 788
- Ippel, M. J., 449, 500
- IQ. *See* intelligence quotient
- Ireland, 1486
- IRT. *See* item response theory
- Irvine, S. H., 583
- Ishioka, T., 533, 559
- ISO. *See* International Organization for Standardization
- Israel-Hamas war, 10, 27
- ISTEP+ test, 1037
- ITC. *See* International Test Commission
- ITED. *See* Iowa Tests of Educational Development
- item analysis: assessment design relation to, 705; of cognitive data summary tables, 1463; IRT and, 702, 703, 704, 718
- item bias, 404, 411
- item bifactor model, 698, 718
- item calibration, of ILSAs, 1451
- item characteristic curve (ICC), 50, 75, 76, 110n8
- Item-Descriptor Matching method, 828, 836
- item development, 151, 460, 489, 722, 996, 1439–40; AIG and, 1292; collaboration for, 994; cut score and, 877–78; item writer and, 511n4; psychometrics and, 471; RPLDs and, 463, 467; TE, 984
- Item Development Checklist Example, 492f
- item difficulty, 850, 861
- item exposure, cheating and, 479
- item factor models, 666, 667, 669
- item features, test-taker characteristics and, 241
- item independence assumption, 351
- item-level adaptive designs, 625
- item-level scoring, 17, 293
- item mapping, 759, 849, 927, 1561
- item parameter estimates, 606, 1533
- item parameter invariance, comparability and, 1510n51
- item pools, norm-referenced interpretation and, 477
- item position effects, 1388
- item preknowledge, 519–20
- item response theory (IRT), 15, 36, 60, 83, 91, 438n9; ability estimates and, 747–48, 780, 798; adaptive testing, 765, 798–99; for BARS, 1375; bias and, 329, 672, 1365; Birnbaum and, 76–77; calibration and, 162, 605, 658, 722, 745, 791, 1548; CAT and, 82–83; classification indices and, 345; complex assessment types and, 338; concurrent calibration, 605; conditional independence and, 671, 1530; cross-classified, 716, 717, 717t, 718, 720, 720t, 721, 721t, 722–23; CTT relation to, 75, 602, 1529, 1530, 1533–34; data collection and, 1359; DCM compared to, 349–51, 668; DIF and, 163–64, 411, 438n8, 679; dimensionality analysis in, 697–98; dominance response process and, 1377; EAP and, 675, 745, 746; equating and, 765, 772, 779–80; error variance and, 316; family-generated clone items and, 749; four-parameter, 112n23; goodness-of-fit tests in, 678; GT and, 357, 1529, 1530; ILSAs and, 1433, 1450, 1451; independence assumption in, 671–72; item analysis and, 702, 703, 704, 718; item characteristic curves and, 50; item independence assumption and, 351; item parameter estimates, 606, 1533; large-scale sampled surveys and, 659; in large-scale testing, 1320, 1548; latent regression model and, 1452, 1534–35; latent trait theory and, 75–76, 1454, 1518, 1531; latent variable models and, 683; learning progressions and, 1066–67; logistic-normal, 673; Lord-Wingersky algorithm and, 696; MAP compared to, 317; measurement error and, 278; MLE and, 745–46; mode-effect models and, 604; multinomial sampling and, 691; NEAT design and, 771; nonlinear transformation of, 331; number-correct scores and, 753; observed score and, 755; performance assessment and, 1388; performance factor analysis and, 628; polytomous, 667, 1421n4; posterior distribution and, 1535; prediction and, 763; proficiency estimates,

- 283–84; proficiency scoring, 322; Rasch model *versus*, 79–80, 412–13, 1560; rater reliability and, 356–57; raters and, 1368; reliability coefficients and, 288, 348, 352–53, 360; reliability in, 280–81, 309–10, 318; RSM and, 1360–61; SAT and, 1204; scale-aligning methods and, 768; scale scores and, 335, 367, 759, 764; scaling and, 70–71, 74–75, 93, 740, 743, 745, 1441–42, 1489; score reporting and, 927; self-report ratings and, 1359; SEM and, 675, 1554; single-group designs and, 603; SJTs and, 1384; speeded assessments and, 354; standard setting relation to, 878, 879; summed score and, 346; TBA and, 600, 602; Thurstonian, 1379–80; two-level, 712t, 713t, 715, 715t, 716t; unidimensional pairwise preference task data and, 1378; validity and, 240; vertical scaling and, 791–92. *See also* multidimensional item response theory; one-parameter logistic model; three-parameter logistic model; two-parameter logistic model; unidimensional IRT
- item slope parameters, in MIRT, 719–20
- item type, cognitive demand relation to, 473
- item writer, 480, 511n4
- item-writing guidelines, 121, 983
- IUA. *See* interpretation/use argument
- Ivancevich, J. M., 524
- Ivy Plus colleges, 1197
- Iwatani, E., 220
- Jackson, C. K., 1141, 1390
- Jackson, D. J. R., 1385–86
- Jackson, J. D., 45
- Jackson, J. J., 1389
- Jackson, K., 1390
- Jacob, B. A., 1139, 1141, 1143–45, 1148, 1149, 1161
- Jacobs, J., 1142
- Jacobsen, E., 1205
- Jacobsen, R., 1134
- Jaeger, R. M., 225, 828, 869, 874, 902, 916, 930, 1312, 1316–17
- James, H. T., 1131
- James, L., 1388
- James, M., 1020, 1022
- Jamshidian, M., 703
- Jankowski, N. A., 1190, 1239, 1241
- Jarjoura, D., 304–5, 328, 781
- Jaschik, S., 400, 407, 1191, 1215, 1263n4
- Jean, M., 350, 366
- Jehangir, 1457
- Jencks, C., 1191
- Jenkins, J., 1053
- Jennings, J. L., 126, 1147, 1148, 1486
- Jennrich, R. I., 703
- Jensema, C. J., 515
- Jensen, Arthur R., 43, 56–57, 352, 416
- Jeon, Minjeong, 17, 673, 681
- Jerald, C. D., 1122
- Jewsbury, P., 158, 587, 601, 603, 610
- Ji, C. S., 1132
- Jia, Y., 615, 794, 1388
- Jo, I., 925
- job analysis, 60, 199, 1281–83, 1289
- job knowledge tests, validity and, 200
- Jodoin, M. G., 352, 362, 984
- Joe, H., 678, 694, 696, 697, 706
- Johannesson, P., 451
- Johannsen, I. M., 946
- Johansson, U., 552
- John, O. P., 1342, 1348, 1350, 1351, 1352, 1356, 1361
- John Birch Society, 7
- Johns, J., 626
- Johns Hopkins, 51
- Johns Hopkins Center for Data-Driven Reform in Education, 1037
- Johnson, C. C., 1061
- Johnson, E. G., 1534–35, 1548, 1561
- Johnson, L., 62, 142, 977
- Johnson, M. S., 348, 365
- Johnson, P. M., 423
- Johnson, R. L., 365, 467, 1029–30
- Johnson, S., 365, 582
- Johnstone, C., 444
- Johnstone, E., 51
- Joint Committee for Guides in Metrology, 1563
- Joint Committee on Testing Practices, 397, 400
- Jones, A. T., 1303
- Jones, L. V., 59, 60, 1119, 1432, 1522, 1533, 1563
- Jones, P., 795
- Jones, S., 1356
- Jonsson, A., 148, 944, 1022, 1027, 1031
- Jöreskog, K. G., 295, 663, 665
- Joshi, M., 552
- Journal of Educational Measurement*, 421
- Journal of Research on Educational Effectiveness*, 659
- Judge, T. A., 1366
- Judson, E., 1147
- Julian, E. R., 1229
- Julian, T., 1210
- Jung, P. G., 1058
- Jung, S., 352
- Junker, B. W., 668, 687
- Jurich, D. P., 1296
- justice, 196, 995–96
- justice-oriented antiracist framework, for validation, 996
- K–12 assessment: accountability and, 194, 1243; assessment development for, 994; Bookmark method in, 872; computer-based assessment in, 498, 1073; cut score in, 226; data story in, 912, 913; diagnostic testing and, 1027; IUA and, 14, 194; norms for, 472; PARCC, 605–6; scale scores in, 755–56, 760; score reporting for, 899, 899f, 900, 930, 943; sociocognitive learning theories in, 243; summative assessments and, 1041; TBA in, 586, 589, 599–600, 607; TE in, 984; test accommodations in, 969; validity theory in, 194, 229, 249–50; vertical scaling in, 789
- K–12 education: accountability in, 21, 483; accountability testing in, 1114, 1115, 1116, 1116f; achievement tests in, 384; ADA relation to, 142; assessment design for, 493; CCSS in, 443; COVID-19 pandemic and, 2; DCM for, 724; ELs in, 973–74; EOC in, 395; ILSAs and, 1425; learning progressions in, 1063; PIRLS and, 1508n20; scaling in, 737; SEL in, 1074; SWDs in, 968
- Kachchaf, R. R., 120, 123, 237
- Kaftandjieva, F., 871
- Kaggle platform, automated scoring and, 557
- Kahneman, D., 126, 834, 838, 841, 842, 843, 844, 845, 847, 860, 871, 893n2
- Kaiaipuni Assessment of Educational Outcomes (KĀEO), 1169
- Kakkonen, T., 547
- Kalantzis, M., 946
- Kaleva, W., 129
- Kaliski, P., 477
- Kallick, B., 1356
- Kallikak, Deborah, 52, 110n9
- The Kallikak Family* (Goddard), 52
- Kameda, M., 533, 559
- Kamens, D. H., 1485
- Kane, M. T., 14, 68, 91, 129, 135–36, 161, 192, 193, 196, 198, 205–12, 213, 215, 216, 219, 221, 222, 224, 225, 226–27, 229, 232, 233, 255, 282, 338, 350, 351, 362, 363, 388–89, 393, 424, 451, 482, 485, 722, 749, 849, 853, 856, 907, 947, 963, 1267, 1268, 1269, 1290, 1297, 1298, 1301, 1303, 1305, 1309, 1310, 1311, 1314, 1315, 1316–17, 1318–19, 1344–45, 1434, 1437, 1476, 1485, 1550
- Kane, T. J., 616, 1143, 1154, 1155
- Kanistra, V., 877
- Kaniuka, T., 1239
- Kankaraš, M., 1351
- Kannan, P., 18, 838, 849, 854, 925, 931
- Kanno, Y., 974
- Kano, Y., 349
- The Kansas Silent Reading Test* (Kelly, F.), 53
- Kaplan, D., 1191–92, 1474
- kappa coefficient, 339
- Kapteyn, A., 1372–73
- Kapur, M., 633
- Karabatsos, G., 742
- Karabenick, S. A., 1479
- Karami, H., 986
- Karantonis, 868
- Karimi, L., 349
- Kasten, N., 1362, 1385

- Katz, I. R., 136–37, 838, 843, 849, 855, 873, 910, 930, 943
- Kautz, T., 1390
- Kay, J., 924
- Kearney, M. S., 1271
- Kearns, J. F., 970, 1064
- Keats, J. A., 327
- Keener, A., 1367
- Keiser, H. N., 1231
- Kell, H. J., 1242, 1343, 1371, 1374, 1375, 1388
- Keller, H., 56, 111n15
- Kelley, T. L., 34, 36, 56, 58, 64–65, 86, 92, 93, 111n18, 112n28, 199, 344, 750–51, 756–57, 762, 763, 764, 769, 792, 1524
- Kelly, F., 53
- Kelly, K., 360
- Kelly, R., 1221
- Kelly v. W. Va. Bd. of Law Exam'rs* (2010), 1285
- Kendall, M. G., 67, 322
- Kendeou, P., 1544
- Keng, L., 194, 242, 249
- Kennedy, A., 1201
- Kennedy, K. J., 117, 138
- Kent, J. D., 1214
- Keppel, Francis, 1131
- Ker, H. W., 1077
- kernel equating, 775, 779, 782
- Kester, L., 829
- Ketterlin-Geller, L. R., 965, 982, 988–89
- Kettler, R. J., 832, 961, 964, 965, 966
- Kevles, D. J., 1522
- Key Reporting Recommendations:
Content—Other Performance Indicators, 940t
- Key Reporting Recommendations:
Content—Report Introduction and Description, 939t
- Key Reporting Recommendations:
Content—Scores and Performance Levels, 939t
- Key Reporting Recommendations: Design, 941t
- Key Reporting Recommendations:
Interpretive Guides and Ancillary Materials, 942t
- Key Reporting Recommendations: Language, 941t
- Key Reporting Recommendations: Overall, 938t
- keystroke logs, 612
- Khalik, S. N., 147
- Khasnabish, S., 946
- Khorramdel, L., 1365, 1452, 1455, 1459, 1477
- Kieffer, M. J., 974
- Kiely, G. L., 82
- Kim, H., 773, 774, 776, 795
- Kim, J. K., 310, 757, 767
- Kim, K. Y., 289, 306
- Kim, S., 314, 317, 318, 319, 351, 548, 747, 795, 796, 799, 1529–30, 1534
- Kim, S. Y., 341, 342, 345, 785
- Kim, Y., 793
- Kim, Y. J., 493
- Kim, Y. K., 754, 789, 791, 792
- kindergarten “readiness” test, 136
- kinesthetic assessment, 552
- King, A. G., 137
- King, G., 1371
- King, M. L., Jr., 1340
- King, T. C., 517
- Kingston, N. M., 601, 849–50, 1031, 1067, 1075
- Kinney, C. L., 1303
- Kintsch, W., 1544
- Kirby, Thomas, 88
- Kirsch, I., 23, 587, 659, 679, 1115, 1432, 1433, 1434, 1436, 1457–58, 1470, 1489, 1510n62
- Kirschenbaum, H., 1044
- Kirschner, P. A., 829
- Kirylo, J. D., 898
- Kishimoto, K., 996
- Kivetz, R., 838
- Klasik, D., 1211
- Klein, S., 1144, 1241
- Klenowski, V., 993
- Klesch, H. S., 1295
- Klieger, D. M., 1215, 1219, 1228, 1371, 1374, 1375
- Klieme, E., 117, 136
- Klingner, J., 147
- Klinkenberg, S., 624
- Kluger, A. N., 1023, 1025
- Klute, M., 1025, 1031
- Knoester, M., 977
- Knollin, H. E., 111n18
- Knowing What Students Know* (NRC), 151, 253, 472, 1551
- knowledge, skills, and other attributes (KSAs), 214, 1345–46; AES and, 255; assessment design and, 219; in classroom assessments, 253; cognitive evidence of, 251; collaborative problem solving assessments and, 259; ECD and, 257–58; for entry-level practitioners, 224; test scores and, 226–27
- knowledge specifications (KS), 709
- knowledge structures, 152, 156–57, 159
- knowledge-tracing technique, 629
- Knox, L., 1206, 1221
- Knupp, T., 931
- Kobrin, J. L., 1226
- Kocher, M. S., 1303
- Koedel, C., 1154–55
- Koedinger, K. R., 467, 1067, 1070, 1074
- Koenig, J., 10
- Koenka, A. C., 1026
- KoKoHs. See *Modeling and Measuring Competencies in Higher Education-Validation and Methodological Innovations*
- Kolb, J. A., 875
- Kolen, M. J., 91, 256, 318, 320, 324, 331, 335, 347, 367, 419, 602, 603, 605, 736, 737, 738, 740, 742–43, 745, 747, 748, 749, 750, 751, 752–53, 754, 757, 758, 759, 760, 765, 766, 768, 769, 770, 771, 772, 773, 774, 775, 776, 778, 779, 780, 781, 782, 783, 785, 789–90, 791, 792–93, 794, 795, 796, 798, 802, 1547, 1558
- Kollias, C., 855, 877, 1388
- Kolstad, A., 1548, 1555
- Konstantopoulos, S., 250, 1037
- Koppich, J. E., 1133
- Kopriva, R. J., 124, 135, 165, 241, 974, 1075
- Koretz, D. M., 21, 130, 802, 1048, 1114, 1117, 1118, 1120, 1122, 1160–61, 1162–63, 1165, 1166
- Koretzian inflation theory, 1160–61, 1162–63
- Korn, S., 1143
- Kosak-Babuder, M., 974
- Kosinski, M., 1390
- Kosslyn, S. M., 945
- KR3, 65
- KR20 coefficient, 65, 66, 111n17, 327; coefficient alpha relation to, 294; reliability and, 364
- KR21, 65
- KR21 coefficient, 294, 343
- Kraft, M. A., 1141, 1142–43, 1145
- Kramer, R. S. S., 552
- Krantz, D. H., 741, 1532, 1577n5
- Krath, J., 530
- Krathwohl D. R., 459
- Kreiter, C. D., 423
- Kridel, C., 1522
- Krippendorff, 357
- Kristof, W., 293, 295
- Kristof's coefficient, 295
- Kroehne, U., 1459
- Kroog, H. I., 1023
- KS. See knowledge specifications
- KSAs. See knowledge, skills, and other attributes
- Kuang, C., 622
- Kubinger, K. D., 1388
- Kuder, G. F., 34, 36, 65, 66, 92, 294
- Kuger, S., 1479
- Kuhfeld, M., 6, 669, 698
- Kūkea Shultz, P., 119, 120, 992, 1169
- Kulhavy, R. W., 1023
- Kulick, E., 411, 412
- Kuncel, N. R., 407, 1227, 1228, 1232, 1366, 1378
- Kwak, M., 548
- Kyllonen, P. C., 23, 387, 583, 962, 1115, 1343, 1345, 1353, 1366, 1373, 1381, 1384, 1479
- Kyngdon, A., 747, 748, 1553, 1561
- Laboratory of Physiology and Psychology, 48
- LaCelle-Peterson, M. W., 973
- Lachlan-Haché, L., 1062, 1142
- Lachman, R., 449
- Ladd, H., 1144
- Lado, M., 1377, 1378

- Ladson-Billings, Gloria, 29n2, 396, 1167
 LaDuca, A., 1290
 Lai, E., 907, 908
 Lai, H., 1292, 1305
 Laing, J., 968
 Laird, R., 1199
 Laitusis, C. C., 968, 969, 1449
 Lam, E. A., 1055
 LaMar, M., 17, 599, 613, 617, 624
 Landauer, T. K., 543, 561
 Landers, R. N., 530
 Landl, E., 877
 Landon, 1304
 Lane, S., 14, 207, 220, 221, 223, 227, 228, 229, 230, 240, 241, 242, 249, 257, 387–88, 442, 443, 448, 470, 511n2, 550, 908, 999, 1018, 1033, 1061, 1114, 1168, 1225, 1301, 1357, 1549
 Lang, K., 1148, 1149
 Langenfeld, T., 401, 402
 language: in assessment design, 237; assessment development relation to, 976; Civil Rights Act relation to, 139; culture and, 13–14, 116, 117, 119–20, 121–22, 975; in score reporting, 936; in sociocultural context, 135; test accommodations for, 517
The Language Police (Ravitch), 395–96
 language testing programs, 193, 219–20, 866; in admissions testing, 386; consequences of, 232, 233–34
 Lapointe, A., 1433
 LaPrade, A., 1238
 large language models (LLMs), 555, 619–20, 637
 large-scale predictive modeling, 656
 large-scale sampled surveys, 659
 large-scale testing, 18, 20, 58, 83; as accountability testing, 1114, 1121; achievement tests, 86; administration conditions in, 124; assessment design of, 120; assessment interpretation of, 194–95; bias in, 121, 133–34, 993–94; classroom assessments relation to, 118, 252, 253; CLR in, 242; cognitive models in, 475; computer-based assessment and, 515–16; consequences of, 230; construct representation in, 130; content referenced growth and, 253; CR in, 245; cut score in, 347; for decision-making, 133–34; diagnostic scoring in, 1018; diversity and, 243; educational disparities in, 979; ELs and, 145; inclusion in, 144; intelligence testing, 52–55; IRT in, 1320, 1548; learning progressions in, 1064, 1068; machine scoring for, 520; MC in, 92; National College Entrance Examination, 89; NCLB and, 1074; New York Regents Examinations, 85; normative information in, 471; PAD for, 445; for personnel selection, 34, 60; PVs and, 1453; reading comprehension and, 1541; RPLDs and, 463; scaling for, 737, 750, 760, 802; score reporting for, 736, 906, 923, 925, 927; sociocognitive learning theories for, 239–40, 1168; sociocultural context in, 118, 128, 128f, 129; subscores in, 920; TBA in, 582; TE in, 984; test scores for, 194; translation procedures for, 122.
See also specific tests
 Larmer, B., 90
 Larrick, R. P., 848
Larry P. v. Riles, 388
 Larsen, M. F., 1135, 1139
 Larson, E. D., 967–68
 latent change variable, regression coefficient and, 716
 latent class analysis, 164
 latent-class signal detection theory model, 357
 latent Dirichlet allocation (LDA), 538, 542–43, 545, 545f, 546–48, 549, 555
 latent growth model, 733n1
 latent proficiency variance, 311, 314
 latent regression, 679–80, 684, 1433, 1453, 1458, 1483; ILSAs and, 1459; IRT and, 1452, 1534–35
 latent semantic analysis (LSA), 543, 544, 547; for AES, 545; dimensionality reduction, 544f; foreign-language scoring engines and, 559
 latent structure analysis, 76
 latent trait theory, 75–76, 1454, 1518, 1531
 latent variable models, 658, 663, 669, 679; Bayesian framework and, 670; data collection and, 679; DCM with, 683, 709; Fisher's identity, 687–88; IRT and, 683; marginal likelihood relation to, 672–74
 Lauen, D. L., 1144–45
 Lave, J., 116, 127, 158, 1544
 Lawley, D. N., 1518
 Lawrence, I., 87, 363, 449, 781, 1202
 Law School Admission Council (LSAC), 1213, 1220–21, 1229
 Law School Admission Test (LSAT), 17, 1191, 1215, 1220–22, 1229, 1232
 Lawton, M., 969
Lawyers for Civil Rights v. President and Fellows of Harvard College (2023), 1201
 Lay, Wilfrid, 45
 Lazarin, M., 517
 Lazarsfeld, P., 60, 75–76, 671, 1525
 Lazarus, S., 969, 970, 981
 Lazowski, R. A., 1348
 LCS. *See* linguistic, cultural, and substantive patterns
 LD. *See* local dependence
 LDA. *See* latent Dirichlet allocation
 L-data. *See* life record data
 Le, Q., 549
 Leacock, C., 538
 learning analytics, 1075–76, 1079
 learning outcomes: accountability for, 136; cognitive ability and, 1238; culture relation to, 126–27; ILSAs and, 1424; sociocultural context relation to, 126–27, 148–49. *See also* student learning outcomes
 learning progressions, 20, 1040, 1052, 1065, 1065f; in assessment design, 1063–64, 1068, 1071; culture relation to, 156; feedback and, 1070; formative assessments and, 1063, 1066–67, 1068, 1069–70, 1071; instructional decisions and, 1062; professional development with, 1069–70
 LeBreton, J. M., 1388
 LeCun, Y., 599
 Leden, L., 148
 Lee, B., 140, 1432, 1468, 1526
 Lee, C. D., 126, 127
 Lee, C. H., 533
 Lee, E., 1031
 Lee, H., 1031
 Lee, J., 1146–47, 1198, 1224, 1354
 Lee, K., 1356
 Lee, M., 1146–47
 Lee, P., 1378, 1379
 Lee, S., 1031
 Lee, T., 698
 Lee, V. E., 393
 Lee, W., 15, 283, 286, 289, 297, 310, 320, 324, 325, 327, 328, 332, 333, 335, 337, 338, 341, 342, 344–45, 346, 347, 360, 361, 363, 749, 778, 783, 789, 796, 803, 1298, 1357, 1367
 Lee, Y., 615, 785, 801
 Lee, Y.-H., 743, 745, 759, 772, 775, 776, 799, 1388
 Lee, Y. J., 126
 Lee, Y. S., 668
 Lefgren, L., 1141, 1155
 Lefstein, A., 139
 Legree, P. J., 1359
 Lehmann, I. J., 995
 Lehrer, R., 1069
 Lei, P.-W., 524
 Leibert, R. M., 127
 Leighton, J. P., 119, 131, 212, 243, 442, 457, 1030, 1292, 1551–52
 Lekholm, A. K., 1043
 Lemann, N., 87, 898, 1191, 1204
 Le Menestrel, S., 139
 Leng, 1384
 Lennon, M. L., 1459, 1489
 Leppo, R., 969
 letters of recommendation, 1366
 Leu, D. J., 1434
 level-specific reliability, 355
 Levenshtein distance, 539
 Levenson, M. L., 29n1
 Leventhal, B., 240
 Levin, H. M., 1340, 1342–43
 Levine, A. I., 1320
 Levine, R., 916
 Levine, R. S., 765, 778
 Levine equating, 769, 778–79
 Levitt, S. D., 1148

- Levy, R., 64, 697–98
 Levy, S., 1320, 1373
 Lewin, T., 1219
 Lewis, C., 82, 338, 343, 615, 1298, 1524
 Lewis, D., 477–78, 480, 722, 829, 849, 877–78, 1547
 Lexile scale, 1560–61
 LGBTQ social movements, 7–8
 Li, D., 767, 1206
 Li, F., 542
 Li, H., 974
 Li, J., 1306
 Li, M., 1030, 1037, 1070
 Li, W., 119, 120, 135, 154, 240
 Li, Z., 694
 LiCalsi, C., 1141
 licensing and certification, 22, 1267, 1269–74, 1279–81, 1289, 1299–1300, 1323; ADA and, 1283–84; assessment development for, 1268; bias and, 1308; credentialing for, 1266; data collection and, 1311; stakeholders for, 1277; test security for, 1286–87; validity and, 1301, 1302, 1303
 licensure, 22, 1266–69, 1323; ADA relation to, 1283–84; Angoff method and, 867, 1314; assessment development for, 1287–89; automated scoring for, 586; certification compared to, 1337n1; culture relation to, 261; cut score and, 1318; IUA and, 1302; in K–12 assessment, 755; linking and, 795; maintenance of competence requirements and, 1277; mandatory professional credentialing and, 1275–76; National Conference of Bar Examiners and, 1279; program sponsors for, 1274–75; regulation of, 1270–73; reliability of, 1297–98, 1300; scale scores in, 760; scaling and, 737; score reporting for, 936; speeded assessments for, 354; stakeholders and, 1313; standard setting for, 1311; test accommodations, 1280; test security for, 1280, 1286–87
 licensure exam, 261, 586, 1281, 1313–14
 Lichtenstein, S., 846
 Lieberman, M., 673
 Lievens, F., 1382, 1383
 life record data (L-data), 1389
 Likert, R., 1358, 1367, 1368, 1372, 1377, 1380–81, 1392
 Likert scale, 1358, 1367, 1377–78, 1380–81, 1479
 Lilja, M., 533
 Linacre, J. M., 356–57, 1368, 1370, 1555
 Linares-Vásquez, M., 944
 Lindquist, E. F., 86, 88, 90, 298, 449, 742–43, 788, 897, 1204–5, 1524
 Lindqvist, E., 1342, 1343, 1366, 1367, 1369
 linear composite, 296, 321
 linear covariance structure models, 698
 linear equating, 765, 769, 775, 780, 781–82
 linear error model, 1520
 linear predictor, 664–65
 linear reduction operator matrices, 693–94
 linear regression model, 416, 556, 663–64
 linear scale alignment, 765, 784–85
 linear transformation, scale scores and, 319
 Lingard, B., 1484
 linguistic, cultural, and substantive patterns (LCS), 12
 linguistic complexity, 119
 linguistic heterogeneity, 119, 135
 linguistic responsiveness. *See* cultural and linguistic responsiveness
 Link, B. G., 138
 linking, 793; adaptive testing and, 772, 798–99; comparability and, 786; data collection and, 738t, 800, 803; equating relation to, 766–67, 769; equity and, 796–97; ILSAs and, 1453–55; NEAT design and, 796, 800; nonequating types of, 787t; PISA and, 794, 1453–54; scale-aligning methods and, 762; score reporting and, 736, 804; sociocultural context and, 16; state-based, 802; subpopulation invariance evaluations and, 783; test administration relation to, 795, 800–801
 linking invariance, 419–20
 Linn, R. L., 90, 91, 151, 220, 221, 227, 230, 240, 384, 389, 403, 417, 598, 600, 765–66, 786, 793, 897, 898, 1126, 1151, 1316–17
 Linquanti, R., 122, 974
 Liou, M., 781, 801
 Lipka, J., 129, 992
 Lipner, R. S., 22, 1291–92, 1294, 1296, 1303, 1304, 1320
 Lipnevich, A. A., 1025, 1046
 Lippmann, Walter, 54–55
 LISREL, 295
 Lissitz, R. W., 455
 literacy, 20, 166, 1468; educational policy for, 1434; ILSAs for, 1427, 1429–30, 1432, 1438. *See also* assessment literacy
 literature and document review, 931
 Little, D., 972
 Little, R. J. A., 679, 1453
 Liu, C., 774, 782, 966
 Liu, J., 409, 415, 420, 438n8, 605, 606–7, 694, 780–81, 782, 783, 786, 788, 1128, 1143
 Liu, K. K., 967, 974
 Liu, O. L., 21, 1072, 1219, 1228, 1237, 1239, 1240, 1241, 1242
 Liu, R., 1067, 1070
 Livingston, S. A., 338, 343, 351, 363, 750, 751, 765–67, 774, 797, 801, 1298
 Livingston and Lewis procedure, 343
 LLMs. *See* large language models
 Local Control and Accountability Plans, 1133
 local dependence (LD), 694, 706
 Lockheed, M., 1434, 1469, 1470
 Lockwood, J. R., 1155
 Lockwood, R. E., 1316
 Loeb, S., 1143, 1144
 Loevinger, J., 200, 202, 203, 204, 275n2, 449
 Loewus, L., 526
 log files, 612–13
 LOGIST, 1534
 logistic-normal IRT model, 673
 logistic substitution, MIRT and, 665
 logistic transformations, 78
 log-linear presmoothing, 773, 781
 Lohbeck, A., 1046
 Lohman, D. F., 449, 457, 500
 Lohr, S. L., 1443
 Lombardo, M. M., 1355, 1356
 Lominger, 1355, 1356
 Longford, N. T., 357
 longitudinal data, 350–51, 1152–53
 Loomis, S. C., 866, 1313, 1317
 Lord, F. M., 60, 70, 72, 74, 76, 77, 79, 80, 82, 90, 112n24, 279, 281, 286, 291, 294, 310, 312, 313, 314, 318, 321, 322, 325, 326–27, 329, 331, 332, 333, 343, 404, 515, 584, 696, 741–42, 743, 744–45, 747, 748, 753, 754, 756, 765, 766–67, 768, 777–78, 781, 784, 796, 797, 879, 1515, 1518, 1525, 1526–27, 1530–31, 1532, 1534, 1535, 1548, 1558–59, 1578n10
 Lord–Wingersky algorithm, 696
 Lorenzo-Seva, U., 317
 Lorge, I., 742, 1522
 Lorie, W., 1159–60
 Lottridge, S., 536, 547, 558, 601, 602, 603, 605, 749
 Louis, T. A., 1456
 Louisiana Department of Education, 444
 Loukina, A., 538
 Loveless, T., 1510n63
 Lovett, B. J., 517
 Lovie, A. D., 42, 43
 Lovie, P., 42, 43
 Lovie, S., 43
 low-income students, 118–19, 136
 LSA. *See* latent semantic analysis
 LSAC. *See* Law School Admission Council
 LSAT. *See* Law School Admission Test
 Lu, R., 134
 Lu, Y., 1308
 Lubke, G. H., 1359
 Lucas, M. E., 1135
 Luce, R. D., 741, 1522, 1532, 1577n5
 Lucido, J. A., 1194
 Ludlow, L., 1371, 1373–74
 Luecht, R. M., 83, 445, 467, 476, 477, 479, 480, 500, 749, 859, 877, 909–10, 1305
 Lui, J. C. S., 520
 Luna-Bazaldua, D., 1431
 Lunz, M. E., 356–57
 Luo, F., 35
 Luo, X., 1478
 Lyell, Charles, 37
 Lyiscott, J., 498
 Lyon, C. J., 1029
 Lyons, S., 1168–69

- Lyons-Thomas, J., 135
 Lytle, S. L., 1071
- MacArthur, C. A., 528
 MacCallum, R. C., 677, 678
 machine-based scoring, 17, 514, 521, 525
 machine learning, 254; cheating and, 520;
 computational psychometrics and, 1075;
 for NGSS, 258; process data for, 615–16;
 response process and, 635; score reporting
 and, 944; topic models for, 542; validity
 of, 538
 machine scoring: data collection for, 526–27;
 for large-scale testing, 520; mark-sense
 readers for, 525; of SAT, 1202; validity of,
 554, 555
- MacIver, R., 908
 MacMillan, D. L., 146
 Madaus, G., 1047, 1048, 1118
 Madiega, T., 622
 Madison, M. J., 350
 Madnani, N., 534, 535, 1366
 Maehler, D. B., 1470, 1510n61
 Maglieri, G., 622
Le magnetisme animal (Binet and Féré), 47–48
 Magouirk, P., 1213
 Mahoney, J. L., 1350
 maintenance of competence requirements,
 1277
 maintenance phase, of score reporting,
 936–37
 mandatory professional credentialing,
 1275–76
 Mangan, D., 27
 Manhattan Institute, 9
 manifest grouping, 164
 Mann, Horace, 36, 84–85, 739
 Manning, C. D., 547
 Manning, W. H., 423
 Mantel, N., 411
 Mantel-Haenszel standardization (MH),
 411–13, 987
 Manyika, J., 137, 1345
 Mao, L., 257
 MAP. *See* maximum a posteriori estimator
 Marais, I., 1533
 Maraun, M. D., 1553
 Marchionni, M., 1474
 Marco, G. L., 403, 788, 789, 1205
 Marcotte, D. E., 1148, 1149
 Marcoulides, K. M., 350
 Marcus, J., 1323
 Mardones-Segovia, C. A., 546
 marginal accuracy index, 341
 marginal agreement index, 341, 344
 marginal category probabilities, 342
 marginal classification indices, 346, 347
 marginal composite scores, 320–21
 marginal fairness, 165
 marginal false-negative error rate, 340, 341
 marginal false-positive error rate, 340, 341
 marginal inference, 124
 marginalized students, 220, 228–29, 497–98
 marginal likelihood, 672–74, 685–86
 marginal maximum likelihood estimation,
 1534–35
 Margolis, M. J., 22, 90, 224, 353, 586, 596,
 851, 870, 962, 992, 1115, 1294, 1308
 Margolis, W., 1213, 1221
 Mari, L., 71, 75, 93, 1515, 1555
 Mariano, L. T., 1141, 1149
 Marianti, S., 614
 Marini, J. P., 518, 1226–27, 1231
 Marion, S. F., 14, 194, 216, 217, 219–20, 231,
 242, 249, 253, 387–88, 443, 483, 511n2,
 550, 908, 998, 999, 1018, 1033, 1060, 1114,
 1127, 1168, 1211, 1225, 1301, 1357, 1549
 Markov chain, 347, 617, 663
 Markov chain Monte Carlo estimation, 347,
 1535
 Markov chain Monte Carlo sampling
 (MCMC), 686, 687, 690, 1379, 1535
 Markov model, 157, 626–27, 627f
 Marks, G. N., 1509n48
 mark-sense readers, 525, 527
 Markus, K. A., 209, 1528, 1531, 1549, 1555
 Marsh, H. W., 1354
 Marsh, J. A., 1034, 1036, 1041, 1050, 1133,
 1136, 1137–38
 Marso, R. N., 1028
 Marston, D., 1055, 1056
 Martens, 1486
 Martin, J., 623
 Martin, M. O., 1441, 1455, 1458, 1471,
 1508n31, 1509n36, 1509n45
 Martin, T., 30n3
 Martineau, J., 998
 Martinez, Y., 127
 Martiniello, M., 120
 Martin-Raugh, M. P., 1388
 Martone, A., 1036, 1156, 1157, 1158, 1159
 Martorell, P., 1141, 1149, 1341
 Maryland School Performance Assessment
 Program (MSPAP), 230–31
 Masci, D., 7
 Mason, B. J., 1025
 Massachusetts Comprehensive Assessment
 System (MCAS), 1037, 1158, 1159
 Massachusetts Institute of Technology, 10
 Massive Open Online Courses (MOOC),
 1238, 1242
 Masters, G., 81, 1378
 mastery learning model, 1119
 matching judgments, 834, 839, 845–47
 mathematical or statistical model, 1524
 Mathematical Tripos, 84
 Mathematics, CCSS, 455
 Mathematics and Reading Frameworks,
 NAEP, 453
 Math Garden, 624, 636
 MATHia, 623, 632, 636, 1073
 Math Mapper 6-8, 1068–69
- matrix sampling design, 1446, 1450
 Mattar, J., 877
 Mattern, K., 21, 1191, 1195, 1198, 1206, 1213,
 1221, 1226, 1231, 1237
 Maul, A., 24, 71, 75, 126, 1517, 1518, 1550,
 1553, 1577n2
 maximum a posteriori estimator (MAP), 309,
 1380; Bayesian estimation and, 317; poste-
 rior distribution and, 1535; reliability and,
 352; for test scores, 674–75
 maximum likelihood estimation (MLE), 309–
 11, 317, 1380, 1452–53; ability estimates
 and, 746, 747–48; bias and, 310, 315; con-
 ditional variance of, 329; IRT and, 745–
 46; marginal, 1534–35; in mixed model
 approach, 357; model-implied summed
 score probability and, 695; proficiency
 scoring and, 322–23; reliability and, 352;
 SEs and, 345–46; variance-ratio defini-
 tion compared to, 316; Variance-Ratio
 Reliability Coefficients for, 315t; vertical
 scaling and, 791
 Maxwell, J. C., 1517–18
 Maxwell, R. S., 516
 Maydeu-Olivares, A., 678, 694, 696, 697, 706,
 1376, 1377, 1379, 1380
 Mayer, A. K., 1236
 Mayfield, E., 533, 558
 Mazzeo, J., 17, 515, 1433, 1437, 1446, 1453–
 54, 1455, 1482–83
 MC. *See* multiple-choice question
 McAbee, S. T., 1362, 1369
 McAllister, P. H., 404
 McArdle, J. J., 712
 McArthur, J., 995
 MCAS. *See* Massachusetts Comprehensive
 Assessment System
 MCAT. *See* Medical College Admission
 Test
 McAuliffe, J. D., 548
 McBee, M. M., 1074
 McCaffrey, D. F., 131, 551, 1155, 1340,
 1367–68, 1370
 McCarrell, T., 88
 McCarthy, M. T., 1214
 McCarty, A. T., 1474
 McClarty, K. L., 851
 McClellan, C. A., 521, 524
 McCloskey, A. V., 978
 McCormick, C. M., 1316, 1317–18
 McCoy, R. G., 1304
 McCrae, R. R., 1342
 McCullagh, P., 664–65
 McCulloch, C. E., 683
 McCusker, M. E., 524
 McDaniel, M. A., 1382, 1383, 1384, 1386
 McDonald, F., 1303
 McDonald, M., 1292
 McDonald, R. P., 349, 719, 1535
 McDonnell, L. M., 978, 1146
 McGaghie, W., 1222

- McGrane, J. A., 24, 71, 75, 1517, 1518, 1522, 1553
- McGrath, R., 499, 501
- McGraw-Hill Education, 623
- McGregor, J., 996
- McIntyre, L., 1518, 1577n2
- McKinsey Global Institute, 1242, 1345
- McLaren, P., 995
- McLaughlin, D. H., 802
- McLaughlin, M. J., 978, 1164
- McMahon, Linda, 26
- McMaster, K., 1055
- MCMC. *See* Markov chain Monte Carlo sampling
- McMillan, J. H., 1035, 1045
- McNamara, C., 1035
- McNamara, D. S., 545
- McNeish, D., 678, 701
- McPeck, W. M., 988
- MCT. *See* minimum competency testing
- McVay, A., 522, 524, 525, 536, 551
- McWhorter, J., 9
- Meachum, J., 1199
- Mead, R. J., 1534
- mean-square error, 785
- Meara, K., 916
- measurement. *See specific topics*
- Measurement, Inc., 554, 555
- Measurement Across the Sciences* (Mari), 71
- Measurement and Prediction* (Stouffer et al.), 60
- measurement error, 69, 279; classification accuracy and, 347–48; equating and, 797; IRT and, 278; in model-based observed scores, 283; psychometrics and, 1554–55; in translation procedures, 1479; true score and, 282
- measurement invariance, 408–10, 424, 1363, 1365, 1370; BARS relation to, 1375; in Big Five personality model, 415; ILSAs and, 1436–37, 1456–57, 1478; linking invariance compared to, 419; prediction invariance compared to, 417–18
- Measurement Research Associates, 927
- measurement theory, 69, 741
- Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe, 1238–39
- measuring scale, 1520–21, 1524
- Medical College Admission Test (MCAT), 404, 1213, 1222, 1229–30, 1232–33
- medical school admissions, test scores relation to, 1214
- Medina, J., 397
- Medina, M. S., 1295
- Medvedev, O. N., 1375
- Mee, J., 841, 851, 1312
- Meeder, B., 624
- Meehl, P., 66, 198, 199, 201, 202, 203, 204, 205, 208, 237, 275n2
- Mehan, H., 158
- Mehrens, W. A., 148, 221, 389, 995
- Meijer, R. R., 162, 425, 519, 1366, 1389, 1549
- Meinck, 1470
- Meltzoff, A. N., 126
- Mendel, G. J., 39
- mental health, COVID-19 pandemic and, 5
- MEPSs. *See* Military Entrance Processing Stations
- meritocracy, 133, 138
- Merriam-Webster, 840
- Mervosh, S., 6, 8, 26
- Messick, S., 13, 24, 34–35, 89, 91, 131, 193, 194, 195, 196, 198, 202, 203–4, 205–6, 207–8, 209–11, 221–22, 228, 232, 233, 236, 388–89, 390, 393, 424–25, 445, 451, 485, 907, 908, 963–64, 1345, 1361, 1425–26, 1433, 1436–37, 1440, 1456, 1466–67, 1475, 1476, 1489, 1491, 1550
- Messick's Facets of Validity, 205f
- metacognition, 1295
- Metropolis, N., 686
- Metropolis-Hastings algorithm, 690–91
- Metropolis-Hastings within Gibbs (Tierney) algorithm, 1379
- Meyer, C., 1061
- Meyer, H.-D., 1425, 1484
- Meyer, J. D., 1345
- Meyer, J. P., 485
- Meyer, R. H., 1359
- MH. *See* Mantel-Haenszel standardization
- MH-RM. *See* Hastings Robbins-Monro algorithm
- Michaelides, M. P., 858, 861
- Michaels, H., 131
- Michaels, S., 1028
- Michel, R. S., 1214, 1215
- Michell, J., 70, 740, 741, 748, 755, 1515–16, 1517, 1518, 1522, 1550, 1552–53, 1557, 1558, 1577n5
- Michigan Assessment Consortium, 1029–30
- Mickelson, R. A., 1144
- Microsoft Corporation, 526, 527
- Middleton, K., 410, 803
- Mikolov, T., 549
- Milbourn, T., 138
- milestone ratings, validity of, 1321
- Military Entrance Processing Stations (MEPSs), 61
- Miller, S., 1037
- Milligan, S. K., 927, 942
- Millman, Jason, 448, 1318
- Mills, C. N., 516, 584
- Millsap, R. E., 409, 410–11, 414–15, 416, 417–18, 1359, 1374
- Milne, R. B., 623
- Milner, A. R., 148
- Mindrila, D. L., 1148
- Miner, B. C., 1043
- minimum competency testing (MCT), 1138–39; consequences of, 1122; *A Nation at Risk* and, 1120–21
- minimum ignorance estimates, 677
- minority populations, 121, 127, 138, 437n6, 1200; NCLB and, 63; placement tests and, 1236; promotion exams relation to, 1150; SAT and, 1231; test accommodations for, 125; test-optional policies and, 1195, 1196–97
- Minstrell, J., 1023
- Minton, H. L., 52
- Mirjafari, S., 1390
- Mirkin, P. K., 1054, 1056–57
- MIRT. *See* multidimensional item response theory
- Mislevy, R. J., 12, 18, 64, 124, 131, 135, 164–65, 166, 195, 209, 236–39, 242–45, 255, 258–59, 260, 316, 399, 442, 445, 456–57, 483, 485, 499–500, 501, 551, 590, 598, 600, 612, 623, 632, 679–80, 739, 740–41, 745, 746–49, 762, 765–66, 793, 803, 909–10, 916, 946, 985, 1072, 1168, 1292, 1305, 1320, 1440, 1452, 1459, 1481, 1515, 1534–35, 1546, 1548, 1550, 1551–52, 1559
- Miura Wayman, M., 1055
- mixed model approach, MLE in, 357
- Miyazaki, I., 384
- Mizumoto, A., 619
- MLE. *See* maximum likelihood estimation
- Mlodinow, L., 131
- M method. *See* Monte Carlo simulation
- MOCA Minute, 1322
- Mock Report of Student Results From Smarter Balanced English Language Arts/Literacy Assessment, 1543f
- mode comparability studies, 601, 603, 605, 794
- mode-effect models, IRT and, 604
- model-based observed scores, 283, 319
- model-based reliability, 349
- model error, 677
- model fit, 79, 662, 694, 698, 723; of DCM, 710; for dimensionality analysis, 676; FP and, 699
- Model Fit Statistics: Single-Level Bifactor Model Versus Two-Level Bifactor Model, 709t
- Model Fit Statistics: Unidimensional Item Response Theory Model Versus Bifactor Model, 708t
- Model Fitting Summary and Key Results: Multidimensional Item Response Theory (IRT) Model Versus Cross-Classified IRT Model, 720t
- model-implied summed score probability, 695
- Modeling and Measuring Competencies in Higher Education—Validation and Methodological Innovations (KoKoHs), 1238
- modeling data, 662
- model tracing, 628
- Modified Angoff method. *See* Angoff method
- Mohadjer, L., 1445, 1510n56

- Molenaar, I. W., 1532
 Molenaar, P. C. M., 162
 Moll, L. C., 993
 Mollenkopf, W. G., 325–26
 Mollenkopf method, 325–26
 Moncaleano, S., 590, 592, 596, 984
Monitoring Educational Equity (NASEM), 1167–68
 monotone nonlinear transformation, 318
 monotonicity property, 740
 Monro, S., 690
 Monroe, A., 1222
 Monroe, S., 694, 707
 Monseur, C., 1480
 Monte Carlo estimation, 347, 677, 690; Bayesian posterior predictive model checking and, 697; latent variable models and, 663; Markov Chain, 686, 687, 690, 1379, 1535
 Monte Carlo simulation (M method), 346
 MOOC. *See* Massive Open Online Courses
 Moon, T. R., 525
 Moore, C. C., 1043
 Moore, D., 841
 Moore, J., 974
 Moors, G., 1358–59
 Morant, G. M., 41
 Morath, E., 1271
 Morell, L., 1069
 Moretti, E., 1341
 Morgan, D. L., 858, 861
 Morgan, J., 1221
 Morillo, D., 1379
 Morris, C. N., 765, 768
 Morris, L. W., 127
 Morrison, K. M., 1305–6
 Morsy, L., 129
 Morton, C., 1056
 Morton, K., 156
 Mosenthal, P. B., 1457–58
 Moses, T., 18, 306, 600, 603, 744, 747, 754, 764, 766, 767, 773–74, 775, 776, 781, 782, 783, 784–85, 789, 795, 799, 801, 802, 803, 1204
 Mosher, F. A., 1062
 Mosier, C. I., 897, 1518
 Moss, P. A., 119, 127, 157–58, 228, 388, 389, 1168
 motivation: CLR for, 498; feedback relation to, 1026; grading relation to, 1046; HSEEs and, 1140; ILSAs relation to, 1480; for SLO, 1240
 motivational equity, 1046
 Motowidlo, S. J., 1384
 Möttus, R., 1373
 Mount, M. K., 1342
 Moustaki, I., 665
 Mozgovoy, M., 521, 542
 Mrkonich, M. J., 137
 Mroch, A. A., 779
 MSPAP. *See* Maryland School Performance Assessment Program
 MST. *See* multistage tests
 Mullainathan, S., 127
 Mullis, I. V. S., 582, 588, 1508n30
 multidimensional item response theory (MIRT), 309, 320–21, 334–35, 671; Bock–Aitkin EM relation to, 689; conditional independence and, 707; cross-classified IRT *versus*, 720, 720t, 721; DCM compared to, 668, 711; factor analysis compared to, 666; item analysis and, 703; item slope parameters in, 719–20; for K–12 education, 724; logistic substitution and, 665; M method for, 347; PV methodology and, 679; score reporting and, 920; vertical scaling and, 792
 multilevel bifactor model, 708, 709, 718
 multilevel construct validity, 217–18
 multimedia, 591, 591f
 multinomial CSEM, 328
 multinomial error model, 333, 344
 multinomial sampling, 691–93
 multiparameter IRT, 79–80, 81, 93
 multiple-choice question (MC), 86, 87, 135, 667, 897; cognition and, 590; in computer-based assessment, 529; for credentialing, 1292–93; with extended-interaction items, 593, 594; extended multivariate GT and, 307, 309; on GMAT Focus Edition, 1220; in large-scale testing, 92; linking and, 795–96; person-by-task interaction and, 163; reliability and, 352–53, 363; RP values in, 759; on Scholastic Aptitude Test, 385; score weights and, 743–44; sociocultural context relation to, 165; on SSAT-II, 397; in TBA, 586, 611–12
 multiple linear regression scoring model, 556
 multistage tests (MST), 351, 352, 353; adaptive testing and, 798–99, 1447, 1482; CAT and, 585–86
 multi-unidimensional pairwise preference (MUPP), 1378
 multivariate generalizability analysis, 70
 multivariate generalizability theory (GT), 304–7, 309, 333–34
 multivariate latent regression model, 680
 multivariate normality, 415, 673
 Mumford, M. D., 1389
 MUPP. *See* multi-unidimensional pairwise preference
 Muraki, E., 81, 667, 1360, 1378, 1534–35, 1548
 Muraki, M., 1360, 1378
 Murphy, G. L., 839
 Murphy, S. C., 1366
 Murray, C., 57
 Musk, E., 25, 27
 Muthén, B. O., 1359, 1379
 Muthén, L. K., 1379
 Mwalimu, 916
 Myers, A. J., 1291
 Myers, C. B., 1039
 Myers-Briggs type Indicator, 1376
 Myford, C. M., 525
 MyLab Writing and Criterion, 636
 Myung, I. J., 698
 Naemi, B. D., 1376
 NAEP. *See* National Assessment of Educational Progress
 Nagaoka, J., 1141, 1149
 NAGB. *See* National Assessment Governing Board
 Nā Lau Lama, 965, 967, 993
 Napolitano, J., 133, 406–7, 1193–94
 narrative patterns, 911–12, 913, 913f, 914, 914f
 Nasca, T. J., 1321
 NASEM. *See* National Academies of Sciences, Engineering, and Medicine
 Nash, B., 1031
 Nasir, N. S., 126
 National Academies of Sciences, Engineering, and Medicine (NASEM), 116, 118, 126–27, 1118, 1132, 1166–67, 1168, 1345, 1350, 1351, 1544
 National Academy of Education, 1131
 National Academy of Science, 1242
 National Assessment Governing Board (NAGB), 442–53, 475, 826, 1354, 1358; equity and, 1167; NAEP relation to, 1131–32; placement tests and, 1234; reading comprehension and, 1544, 1565, 1566
 National Assessment of Educational Progress (NAEP), 6, 26, 442, 975, 1039, 1450; accountability and, 21, 63, 93, 1147–48, 1340; accountability testing and, 1115; achievement gaps and, 1051; BARS and, 1458; data collection and, 679, 1354; Data Explorer, 923, 1464–65; data story for, 912; DIF and, 413; educational disparities and, 979; equity and, 1167; Grade 4 Mathematics Policy-Level Definitions, 461f; group-score assessments, 601; ILSAs and, 1424, 1432–33, 1489–90; impact data and, 864; IRT and, 1548; large-scale sampled surveys and, 659; linking and, 794–95, 801–2; Mathematics and Reading Frameworks, 453; motivation and, 1480; NAGB relation to, 1131–32; National Institute of Education and, 89; *A Nation at Risk* on, 1120; norms and, 758; policy makers and, 1475; policy PLDs and, 460; reading comprehension and, 1541, 1542, 1544, 1545f, 1546, 1556f, 1558f, 1565, 1565t, 1566; scale scores and, 759, 1556, 1557, 1558–59; scaling for, 760, 1547, 1561; score reporting and, 902, 916, 927, 1119–20, 1163; self-report ratings and, 1358; short-answer scoring and, 538; TBA and, 587–88, 599, 603–4, 607–8, 611; Technology and Engineering Literacy assessment, 475, 476, 587, 594, 598; TEIs and, 587–88; test-based inflation and, 1161

- National Association of Colleges and Employers, 1242, 1344
- National Association of Securities Dealers, 584–85
- National Association of Teachers of Educational Measurements, 1577n3
- National Benchmark Tests, in South Africa, 873
- National Board for Professional Teaching Standards, 500, 983
- National Board of Medical Examiners (NBME), 586, 596, 1285–86
- National Center for Educational Outcomes, 394, 495, 973–74, 981
- National Center for Education Evaluation and Regional Assistance, IES, 1134
- National Center for Education Statistics (NCES), 393, 401, 555, 587, 927, 974, 980, 1132, 1263n1, 1263n8, 1354, 1366, 1507nn11–12
- National Center for Homeless Education, 127
- National Center for Learning Disabilities, 1129
- National Center on Education and the Economy, 1121
- National Center on Response to Intervention, 1053
- National College Entrance Examination, 90
- National Collegiate Athletics Association, 1210
- National Commission for Certifying Agencies, 905, 1300
- National Commission on Certification of Physician Assistants, American Board of Internal Medicine, 1274
- National Commission on Excellence in Education, 898, 1120
- National Conference of Bar Examiners (NCBE), 1279, 1286–87
- National Conference of State Legislatures, 1140
- National Council of Architectural Registration Boards (NCARB), 586, 1278–79, 1287–88
- National Council of Examiners for Engineering and Surveying, 1294
- National Council of Teachers of Mathematics, 1121
- National Council on Disability, 981
- National Council on Education Standards and Testing, 1121
- National Council on Measurement in Education (NCME), 216, 389–90, 437n5, 975, 997–98, 1130, 1190, 1515, 1546, 1564, 1577n3
- National Defense Education Act (NDEA), 61–62
- National Educational Longitudinal Study, 1366, 1370
- National Education Longitudinal Study, 974
- National Governors Association, 1131
- National Institute for Learning Outcomes Assessment (NILOA), 1237
- National Institute of Education, 89
- National Intelligence Tests, 386
- National Literacy Project, 1468
- National Longitudinal Study of Youth, 1387
- national norms, 758
- National Research Council (NRC), 59, 119, 126, 141, 146, 148, 149, 151, 210, 244, 252, 442, 446, 454, 457, 472, 801, 1062, 1143, 1145, 1342–43, 1344–45, 1350, 1544, 1551
- National Science Foundation, 1119
- National Society of Professional Engineers, 1266
- National Student Clearinghouse Research Center, 1241, 1390
- National Teacher Examination, 869
- A Nation at Risk* (report), 62, 898, 1120–21
- Nation's Report Card*, 1340
- native language skills, 976
- Natriello, G., 1020
- Natural Inheritance* (Galton), 40
- natural language processing (NLP), 131, 533, 549, 558; for diagnostic feedback, 535; foreign-language scoring engines and, 559
- NBME. *See* National Board of Medical Examiners
- NCARB. *See* National Council of Architectural Registration Boards
- NCBE. *See* National Conference of Bar Examiners
- NCBE v. Multistate Legal Studies* (2006), 1286–87
- NCES. *See* National Center for Education Statistics
- NCLB. *See* No Child Left Behind Act
- NCME. *See* National Council on Measurement in Education
- NDEA. *See* National Defense Education Act
- Neal, D., 1148, 1163
- NEAT design. *See* nonequivalent groups with anchor test
- Nedelsky, L., 838
- Nedelsky method, 838
- Neilsen, N., 10
- Nelder, J. A., 664–65
- Nelson, J., 970
- Nelson, J. M., 517, 969
- Nelson, L. S., 1303
- Nelson-Barber, S., 119, 131, 140, 149, 154, 236, 992
- Nering, M., 1421n4, 1549
- The Nested and Hierarchical Nature of Claims, 488f
- Nester, M. A., 972
- Neubauer, A. C., 1385
- neural networks, 628, 749
- Neustel, S., 1289
- New Data Show How the Pandemic Affected Learning Across Whole Communities* (report), 4–5
- “A New Design for a New Era,” 1433
- New Jersey School for Feeble-Minded Boys and Girls, 51–52
- New Republic* (Lippmann), 55
- New Standards Project, 1121–22
- Newstead, S. E., 1359
- Newton, P. E., 206, 211, 222, 224, 1049, 1549, 1550
- Newton–Raphson technique, 1534
- New York Regents Examinations, 85
- New York State Testing Program (NYSTP), 461, 463, 466
- next-generation assessments, score reporting for, 946
- Next Generation Science Standards (NGSS), 7, 442–43, 454, 480, 494, 500; AP and, 446; Inq-ITS and, 624; machine learning for, 258
- Nexus Capital Management LP, 1206
- Neyman, J., 663
- Ng, H. L., 1468
- n-gram vectors, 555
- NGSS. *See* Next Generation Science Standards
- Nguyen, N. T., 1383
- Nicewander, W. A., 310, 352
- Nichols, P., 16, 152, 445, 449, 457, 460, 485, 486, 500, 840
- Nickerson, R. S., 838
- Nickodem, K., 979
- Niepokoj, D., 223
- Niessen, A. S. M., 425, 1362, 1366, 1389
- Nikolic, S., 137
- NILOA. *See* National Institute for Learning Outcomes Assessment
- Nilson, T., 1475
- Nisbet, I., 194
- NJ Smart, 1390
- NLP. *See* natural language processing
- Noble, J., 1225, 1226, 1231
- Noble, S. U., 137
- No Child Left Behind Act (NCLB), 21, 62, 133, 483, 898, 977; accountability and, 1049–50, 1118, 1123–27, 1146, 1151, 1156, 1340; accountability testing and, 1114; alignment and, 1159, 1160; culture relation to, 140; cut score and, 866; ELs and, 1164–65; EOC and, 1139; ESSA compared to, 1129; interim assessment relation to, 250, 1032, 1074; large-scale testing and, 1074; NAEP relation to, 1131; noncognitive skills and, 1354–55; performance standards in, 878–79, 1122; racialized status relation to, 977–78; report cards and, 1132–33; school performance relation to, 1148; SLOs and, 1059, 1060; sociocultural context and, 134; state summative assessments and, 1047; state testing and, 219; student achievement relation to, 1143–44; SWDs and, 63, 143–44; test accommodations and, 142–43

- Noem, Kristi, 26
 Noffle, E. E., 1342, 1343, 1362
 nominal response model (NRM), 80, 1359, 1360, 1384
 nomological network, 201–2, 237
 nomological validity, 203
 noncognitive skills, 1341–42, 1344–45, 1347, 1392; cognitive ability relation to, 1390–91; NCLB and, 1354–55; performance assessment for, 1387–88; personality traits and, 23, 1343
 nonequating types, of linking, 787*t*
 nonequivalent groups, 604–5
 nonequivalent groups with anchor test (NEAT design), 765, 769, 775, 785; for data collection, 771; equating and, 776–77, 778; linking and, 796, 800; vertical scaling and, 790
 nonlinear performance trajectories, 1125
 nonlinear structural equation model, 349
 nonlinear transformation, 319, 754; of CSEM, 329–30; of IRT, 331; scale scores and, 752–53
 nonresponse bias, 1445
 Noonan, L. E., 524
 NORC, 1159
 Norcini, J. J., 1303–4, 1314, 1319
 normal approximation method, 341–43
 normality assumption, 345, 670–71
 normal ogive, 75, 76, 1530–31
 normal theory maximum likelihood, 676
 normative feedback, 850–51
 normative information, in large-scale testing, 471
 norm-referenced interpretation, 50; interim assessment and, 1036; item pools and, 477; percentile-rank system, 479
 norm-referenced testing, 897, 898; for admissions testing, 1210; criterion-referenced tests compared to, 1118–19; score reporting and, 919
 norms: auxiliary scales and, 757–58; DIF and, 479; formative assessments relation to, 1030; for K–12 assessment, 472
 North Carolina Board of Dental Examiners, 1280
 Northwest Evaluation Association, 631, 1033, 1037
 Norton, L., 1046
 Nortvedt, G. A., 195, 236
 Novakovic, N., 871, 873, 874
 Novarese, M., 1390
 Novick, M. R., 60, 90, 281, 286, 291, 294, 338, 343, 408–9, 421, 422, 741–42, 743, 744–45, 753, 754, 767, 784, 1515, 1524–25, 1526–27, 1530–31, 1548, 1558–59
 NRC. *See* National Research Council
 NRM. *See* nominal response model
 number-correct scores: coefficient alpha and, 280; IRT and, 753; reliability and, 352, 363; TBA and, 600
 Nungester, R. J., 1300–1301, 1318
 Nunnally, J. C., 920, 1223, 1360
 Nye, C. D., 1365
 NYSTP. *See* New York State Testing Program
 Obama, B., 192
 Oberski, D. L., 349
 Objective Performance Index, 920
 Obradovic, J., 127
 observable evidence, 452, 458–59, 469, 596
 observation model, 151, 154, 155*f*, 155–57, 167
 observed score, 278, 417, 755; model-based, 283, 319; reliability and, 281, 1061
 occupational overshift, 22
 O'Connell, M. S., 1386, 1509*n*48
 O'Day, 1121, 1156
 odd-even split, 293
 O'Donnell, F., 19, 872, 905, 915, 925, 928
 O'Donnell, S., 1467, 1468, 1472–73, 1484
 O'Dwyer, L. M., 1475
 OECD. *See* Organisation for Economic Co-operation and Development
 Office of Naval Research, 60
 Office of Personnel Management, U.S., 1344, 1382
 Office of Technology Assessment, 149, 385, 386
 Office of the President of University of California, 1193
 official language policies, 135
 Ogasawara, H., 360, 781
 Oh, I.-S., 1227, 1368, 1369
 Ohland, M. W., 1375
 Ohm, George, 112*n*22
 OIB. *See* ordered item book
 Okada, K., 1380
 Ola, B. A., 127
 Oláh, L. N., 1035, 1036
 Oldfather, P., 1042
 O'Leary, T. M., 210, 908, 947
 Oliva-Córdova, L. M., 928
 Oliveri, M. E., 119, 123, 135, 162, 164, 240, 259, 983, 986, 1484, 1486
 Olkin, I., 1119, 1432
 Olsen, B., 1045
 Olsen, L. W., 77
 Olson, E., 1215
 Olson, J. F., 969
 Olson, L., 587, 1130
 O'Malley, K., 851
 1PL model. *See* one-parameter logistic model
 O'Neill, K. A., 987–88
 O'Neill, T. R., 1308
 one-parameter logistic model (1PL model), 112*n*26; ability estimates from, 746; Rasch model and, 1532; scaling and, 745
 Ones, D. S., 1368, 1369, 1370
 online environments, 136, 1307–8
On the Correlation of Mental and Motor Ability in School Children (Bagley), 42
On the Origin of Species (Darwin), 38
On the Tendency of Varieties to Depart Indefinitely From the Original Type (Wallace, A.), 37–38
 ontoepistemic heterogeneity, 152
 on-track labels, educational expectations and, 207
 Oostrom, J. K., 1382, 1386
 open-admission institutions, 1263*n*8
 operating motivations, 130
Operational Best Practices for Statewide Large-Scale Assessment Programs, 448
 Operational Personality Questionnaire (OPQ), 1376
 operational testing programs, 216, 466, 585, 1315
 Oppen, I. M., 1077
 opportunity-to-learn (OTL), 10, 398, 478, 479
 OPQ. *See* Operational Personality Questionnaire
 optical scanner, 34
 opt-out movement, 898
 oral examinations, 84, 1302
 oral responses, 528
 Oranje, A., 131, 254–55, 587, 1452, 1548, 1555
 ordered item book (OIB), 843, 844, 849, 867–68, 894*n*9
 order effects, 603
 O'Reilly, T., 594
 Organisation for Economic Co-operation and Development (OECD), 139–40, 393, 588, 592, 596, 604, 622, 1238, 1340, 1351–52, 1354, 1356, 1383, 1388, 1427, 1430, 1432, 1434, 1436–37, 1444, 1449, 1461, 1466, 1471, 1484, 1486
 organizational psychology, 1389
 Organization and Construct Coverage of e-rater Version 19.1, Released in 2019, 537*f*
 O'Riordan, M., 610
 Orlando, M., 317, 694, 696, 742, 747
 Ormerod, C., 538
 orthonormal matrices, 543, 544
 Ortner, T., 1388
 Osborn, F., 59
 Osborne, M. A., 1345
 Ostini, R., 1421*n*4
 Östling, R., 533
 O'Sullivan, B. J., 524
 Oswald, F. L., 1345, 1346, 1355, 1362, 1363, 1365, 1374, 1377, 1383, 1385, 1386, 1389
 Otero, V. K., 1023, 1029
 Otis, Arthur, 54, 111*n*18, 764
 Otis advanced mental test, 45
 Otis Group Intelligence Scale, 54
 Otis Self-Administering Tests of Mental Ability, 54
 OTL. *See* opportunity-to-learn
 O'Toole, B. I., 1421*n*1
 Otterman, S., 26
 Ou, D., 1148

- outcome feedback, 1025, 1026
 outcomes assessment, 20–21, 1240–41, 1271
 Overbaugh, R., 134
 overconfidence bias, 846
Overview of Assessment Contexts for Reporting, per the Standards for Educational and Psychological Testing, 904f
 An Overview of More Specialized Models in the General Framework, 661t
 Owens, K., 129
 Owens, W. A., 1389
 Owings, J., 393
 Özek, U., 1141, 1149–50
 Ozer, D., 1342–43
- Paap, M. C. S., 583
 Packman, S., 1237
 PAD. *See* principled assessment design
 Padilla, J.-L., 614, 945
 Paek, P., 601
 Pérez, D., 127
 Page, E., 550, 561
 Pager, T., 27
 Pal, S., 616
 Palestine, 10, 27
 Palmer, 232, 233
Palmer College of Chiropractic v. Davenport Civil Rights Commission (2014), 1284
 Panadero, E., 1022, 1027, 1031
 Pane, J. F., 623, 630
 panel diversity, 852–54
 panelists, 828, 871, 894n8, 936; Angoff method relation to, 865–67, 872–73; Bookmark method relation to, 867–68, 871; borderline test takers relation to, 869; cognitive and social model relation to, 873–74; cognitive-judgmental task and, 833–34, 841, 865, 876–77; content expertise relation to, 870; facilitator and, 17; in group interaction submodel, 831–32; NCLB relation to, 878–79; probability judgments and, 838, 844; SES of, 870; for standard setting, 1313–15; standard-setting design relation to, 850, 852; threshold-signal detection judgments and, 840; training of, 860–61, 1315; in workshop process submodel, 858–59
 Paolucci, C., 1475
 Papa, R., 995
 Papadakis, M. A., 1303
 Papageorgiou, S., 874
 Papay, J. P., 1148, 1155
 paper-and-pencil (PBA), 604, 610; ILSAs and, 1446, 1477; linking and, 794–95; matrix sampling design and, 1450; NAEP and, 608; PARCC and, 605; TBA compared to, 599, 600–603, 607
 para-data, 1445
 parallel-forms coefficient, 313–16
 parallelism, 296, 300; part-test, 291
 parallel-linear linkage models, 420
- PARCC. *See* Partnership for the Assessment of Readiness for College and Career
 Parent Questionnaire, PISA, 1366
 parents, accountability relation to, 1133–35
 Paris, D., 497, 1167
 Park, R., 352
 Park, Y., 925
 Parke, C. S., 230
 Parker Webster, J., 992
 Parsad, B., 1234, 1236
 Parshall, C. G., 528, 1293–94
 Parsi, A., 1047
 partial correlation, 43–44
 partial credit model (PCM), 81, 357, 412, 1360–61, 1451
 Partnership for the Assessment of Readiness for College and Career (PARCC), 461, 463, 556–57, 587, 599–600, 605–6, 608, 1158
 part-test parallelism, 291
 part-test variance, congeneric forms and, 295–96
 Pashler, H., 698
 Pashley, P. J., 423
 Pate, D. T., 26
 A Path Diagram Depicting the Level 1 (Individual-Level) Model, 714f
 A Path Diagram Depicting the Level 2 (Site-Level) Model, 714f
 Patil, A., 402
 Patterson, B. F., 1226, 1231
 Patterson, C. H., 366
 Patterson, F., 1382
 Patz, R. J., 357, 525, 687, 1296
 Paulhus, D. L., 1362
 Paunonen, S. V., 1352
 Pavlik, P. I., Jr., 628
 Pavlov, G., 1378
 PBA. *See* paper-and-pencil
 PCM. *See* partial credit model
 PD. *See* productive dichotomous items
 Peabody, M. R., 1303
 Pearlman, M., 445, 451, 500
 Pearson, K., 34, 40–42, 43, 44–45, 64, 73, 84, 92, 93, 110n3, 358, 540, 606
 Pearson correlation, 313, 1558
 Pearson product-moment correlation, 557
 Pearson *r* correlation, 556
 Pearson VUE, 516
 Pedder, D., 1020
 Pedulla, J. J., 1047, 1049, 1146
 peer assessors, 1030
 Peeters, H., 1383
 PEG. *See* Project Essay Grade
 Peiser, B. M., 1134–35
 Pellegrino, J. W., 131, 212, 217, 219, 239, 243, 251, 252, 254, 442, 454, 500, 613–14, 632, 749, 759, 823, 848–49, 927–28, 1066, 1072, 1238, 1481
 Penev, S., 349
 Penfield, R. D., 410, 411, 412, 1150
- Peng, C. J., 341, 342, 343
 Penner, E. K., 1168
 Pennington, C. R., 154
 Pennock-Roman, M., 610
 Penuel, W. R., 237, 252, 442, 444, 986, 1033, 1034
 Pepper, D., 119, 1437, 1481
 Percentage of 4-Year Institutions by Undergraduate Admission Rates, 1208t
 Percentage of U.S. Applications and Enrollment by Selectivity at 4-Year Institutions for 2017–2018, 1209t
 percentile-rank system, 472, 479
 Pereira, M., 1034
 Perez, C., 88
 Perfetti, C., 1544
 performance assessment, 134, 220, 352–53, 1386; assessment literacy and, 1040; common formative assessments as, 1038; consequences of, 230–31; construct-irrelevant variance in, 1388–89; for non-cognitive skills, 1387–88; performance standards in, 1121–22; sociocultural context relation to, 165
 performance-based instruction, 230, 452
 performance categories, psychometrics for, 447
 performance factor analysis, adaptive learning and, 628
 performance inferences, 212–13
 performance in practice, credentialing relation to, 1304
 performance labels, 905
 Performance-Label Word Cloud, 905f
 performance-level descriptors (PLDs), 447, 839, 841, 847, 850; in Angoff method, 866–67; in assessment design, 877, 878; in Bookmark method, 867; cognition and, 457, 458–59; cognitive complexity and, 842; cognitive-judgmental task and, 829, 859; domain analysis and, 456; ID Matching and, 844–45; in PAD, 452–53, 460; panelists relation to, 872–73; performance standards and, 832; policy, 460, 872; target, 464, 465; validation for, 489. *See also* range PLDs
 performance standards, 207, 226–28, 1310; for alternate assessments, 970; consequences of, 224–25; for credentialing, 822; cut score and, 474, 832; in NCLB, 878–79, 1122; in performance assessment, 1121–22; PLDs and, 832; score reporting and, 736–37
 Perie, M., 219, 250, 454, 458, 476, 483, 859, 1032, 1033
 Perin, D., 561
 Perjons, E., 451
 Perkins, P., 552
 Perline, R., 742, 746, 747, 1533, 1553
 permissible statistics, 1521–22
 Perrenoud, P., 1022

- Perrote, A. J., 548
 personal feedback, 1026
 personality tests, 201, 206, 414; Big Five
 personality model and, 415, 1356; scaling
 and, 1357
 personality traits, 23, 1343, 1346
 personalized learning systems, 623–24
 Personal Potential Index, 1367–68, 1370
 person-by-task interaction, 162–63, 598
 person fit, 162, 519
Personnel Administrator of Mass. v. Feeney, 149
 personnel selection, large-scale testing for,
 34, 60
 persuasive argument theory, 894n8
 Peters, R., 1042
 Petersen, N. S., 421, 422, 736, 737, 738, 742–
 43, 749, 752, 758, 760, 762, 765, 766, 770,
 772, 797, 802
 Peterson, C., 1356
 Peterson, J. J., 127, 1350
 Peterson, P., 1071
 Petrilli, M. J., 1128
 Petrusel, R., 493
 Petscher, Y., 354
 Pham, D. N., 1064, 1066
 Phelps, R. P., 85, 130
 Phillips, M., 1137, 1138, 1191
 Phillips, S. E., 149, 402–3, 404, 405, 406, 408,
 969, 1273, 1279, 1280, 1281, 1284
 philosophy, of science, 24
 physical stimuli, 71–72
 PIAAC. *See* Programme for the Assessment of
 Adult Competencies
 Pianta, R. C., 1128
 Pickford, P., 1046
 Piech, C., 628
 Pigge, F. L., 1028
 PIRLS. *See* Progress in International Reading
 Literacy Study
 PISA. *See* Programme for International
 Student Assessment
 PISA 2015 Extended-Interaction Item From
 the Running in *Hot Weather* Set, 593f
 PISA 2015 Interactive Science Item in Which
 Students Use Drag-and-Drop Interactions
 to Complete a Model of a Fish Farm, 592f
 Pitoniak, J., 29n1
 Pitoniak, M., 224, 226, 756, 828, 852, 856,
 860, 868, 873, 968, 1191–92, 1310, 1313,
 1318–19, 1547
 Pitt, E., 1046
 Pitt, M. A., 698
 placement tests, 361, 795, 1190, 1233, 1235t,
 1243; ACCUPLACER, 1234, 1236–37;
 validity of, 1234, 1236
 plagiarism detection, 520–21
 Plake, B. A., 1311
 Plake, B. S., 390, 476, 477, 480, 489, 500, 849,
 851, 853, 859, 860, 865, 870, 874, 1313
 plausible values (PVs), 1448, 1453, 1457, 1458,
 1464, 1509n39
 PLDs. *See* performance-level descriptors
 Plous, S., 838, 841, 847
 Ployhart, R. E., 1363, 1365, 1385
 P method. *See* individual proficiency estimates
 Poe, M., 995
 Poisson model, 78, 546
 Pokorny, H., 1046
 Pokropek, A., 1474, 1480
 polarizing topics, as construct-irrelevant
 variance, 395–96
 policy assumption, 1316
 policy decisions, 226, 999, 1116; account-
 ability data for, 1137–38; standard setting
 and, 1316, 1317–18
 policy makers, 132, 148; ILSAs relation to,
 1470, 1473, 1485–86, 1489; NAEP and,
 1475
 policymaking, standard setting as, 820n1,
 822, 851
 policy PLDs, 460, 872
 Policy Research, ILSAs and, 1489
 Polikoff, M., 21, 148, 227, 386, 483, 586, 658,
 659, 962, 977, 1116, 1117, 1122, 1123,
 1124, 1125, 1126, 1130, 1132, 1134, 1143,
 1145, 1147, 1156, 1157, 1158, 1159, 1190,
 1211, 1390
 political/cultural climate, 4, 7–8, 13
 Pollio, M., 1043
 Polson, C., 1148
 polynomial regression coefficients, 326, 334
 polytomous IRT models, 667, 1421n4
 Pommerich, M., 610, 762, 763, 786, 788, 794
 Poniatowski, P. A., 1290
 Pons, X., 1468, 1470, 1486
 Pools, E., 1480
 Poortinga, Y. H., 147
 Popham, W. J., 126, 151, 221, 389, 979, 1298
 Popperian philosophy, 699
 population covariance matrix, 677
 population heterogeneity, 163–64, 167, 196
 population homogeneity, 121, 122
 population invariance, 419, 765–69, 777–78,
 780, 782–83, 802
 population model, 1452, 1483, 1549
 population sampling, 1443–45
 Por, H. H., 119, 154, 165
 Poropat, A. E., 1342, 1343, 1362, 1368–69
 Porter, A. C., 533, 1018, 1122, 1123, 1125,
 1147, 1157
 portfolio Assessment, 1047, 1122
 Posselt, J. R., 1214–15
 Posterior Distribution, MLE and, 1535
 post hoc alignment, 493
 Potts-Datema, W., 127
 poverty, 131, 393, 979, 980
 Powell, D., 1051
 Powell, S., 1315, 1317, 1319
 power asymmetry, social inequities and, 138
 Powers, D. E., 255, 619, 1228
 Powers, S., 878
 PP. *See* productive polytomous items
 practice analysis, 224, 226, 1283, 1288,
 1289–91, 1303
 practice responses, 523
A Practitioner's Guide to Growth Models
 (Castellano and Ho), 1152
A Pragmatic Perspective of Measurement
 (Irribarra), 71
 Prasad, J., 1386
 Praxis test, scale scores for, 750, 751
 Preacher, K. J., 698
 PreACT, 1205
 precision, 364–65, 366
 precision estimates, 360
 prediction, 91, 763, 784–85, 793
 prediction invariance, 409, 415–19, 438n8
 predictive bias, 388, 417
 predictive validity, 201, 240, 1034, 1043,
 1471; CBM and, 1056; GMAT and, 1232;
 GRE and, 1228; MCAT and, 1229, 1233;
 SAT and, 1226
 Prediger, D. J., 1298
 pre-equating, 765–66, 771–72, 776, 798
 preknowledge, 519, 1306–7
 Preliminary Multistate Bar Review, 1286–87
 Prendergast, C. O., 582
 pre-service teachers, 454
 Pressey, S. L., 622
 Preston, J. P., 993
 Preston, K., 1361
 PReview, 1222
 Price, B., 1344
 primary scales, 738, 759, 760
 Primi, R., 1365, 1373
 Princiotta, D., 1146
 Principal Component Analysis of
 21st-Century Skills and Big Five
 Alignment, 1349t
 principal evaluation, Bookmark method for,
 870
 principled assessment design (PAD), 445–
 46, 453, 474, 487t, 501, 511n1; accessi-
 bility and, 473, 494–95, 982; assessment
 constraints and, 459–60; assessment
 development and, 452, 472–73, 983; CLR
 in, 497–98; cognition and, 475; cognitive
 complexity and, 469, 470; design pat-
 terns in, 465, 466, 467; documentation
 and, 447, 451–52, 489, 497, 910; domain
 modeling in, 456; ECD and, 448, 452,
 500; evidentiary argument in, 482–83,
 485–86, 487, 493; procedural validity
 evidence in, 489; psychometrics and, 449,
 450, 471–72, 473; range PLDs (RPLDs)
 and, 476–78, 480, 492; reasoning and,
 447; RPLDs and, 461, 476–78, 480,
 492; score interpretation in, 451; score
 reporting and, 909–11; UDA and, 982,
 994; UX and, 499; validation and, 451,
 481–82; validity and, 470, 494
 Principled Assessment Design Essential
 Elements, 453f

- principled item development process, cut score in, 877–78
- Principles for the Validation and Use of Personnel Selection Procedures*, 1282
- Prior, P., 553–54
- privacy, 402, 405, 633
- proactive recipience, for feedback, 1027
- probabilistic modeling, 623, 624, 637
- Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch), 77–78
- probability judgments, 834, 837, 844, 845, 847; Angoff, 848–49, 850; bias in, 838, 846
- probit link function, 665
- problem-solving process, 252, 594, 595, 596, 597
- procedural validity evidence, 482, 486, 487, 489
- process data, 613, 614–16, 660, 942–43
- process feedback, 850
- productive dichotomous items (PD), 719–20
- productive modality, 718, 719
- productive polytomous items (PP), 719–20
- Professional Certification Coalition, 1322
- professional degree admissions, 1213–15
- professional development, 147, 483, 484, 485, 659, 1040; CBM and, 1057, 1059; certification and, 1321–22; formative assessments and, 1029–30, 1039; grading and, 1045; with learning progressions, 1069–70; teacher resource material for, 1036–37
- professional judgment, validation framework and, 199
- professional practice model, 1290
- Professional Testing Corporation, 1311
- proficiency estimates, 283–84, 351, 353, 354, 626
- proficiency scoring, 322–23
- program accreditation, 1190, 1271, 1274
- program evaluation, 62, 141, 1118, 1369, 1541; CBM for, 1053; validation framework and, 208
- Programmed Logic for Automatic Teacher Operations system, 584–85, 589
- Programme for International Student Assessment (PISA), 26, 136, 138, 386, 1340, 1351; accountability testing and, 1115; adaptive testing and, 1447–48; AVs and, 1373; background questionnaire relation to, 1442; bias in, 1472; cognitive scales for, 1457–58; comparability and, 1485; data collection and, 679, 1474; DIF and, 240; educational policy relation to, 1468–69, 1486, 1490; education reform movements and, 139; forced-choice methods and, 1381; IDB Analyzer and, 1470; IRT and, 1451; linking and, 794, 1453–54; literacy and, 1427, 1429–30, 1432, 1438; noncognitive skills in, 1354; Parent Questionnaire, 1366; population sampling and, 1444; process data from, 614; PUF and, 1462; PVs and, 1458, 1509n39; self-report ratings and, 1358; SJTs and, 1383; SLO and, 1240; survey congestion and, 1480; TBA and, 587, 592, 593, 594, 596, 603, 604; test accommodations and, 1449; test administration for, 1478; validity and, 1437
- Programme for the Assessment of Adult Competencies (PIAAC), 1351, 1430, 1444, 1464, 1465, 1510n57; adaptive testing and, 1447–48; background questionnaire and, 1431–32, 1434; bias in, 1456; cognitive scales for, 1457–58; data collection and, 1471, 1474, 1488; as digital assessment, 1483; educational policy relation to, 1469; gender and, 1471; literacy and, 1438; para-data and, 1445; PUF and, 1462; survey congestion and, 1480; translation procedures for, 1442
- program sponsors, 1274–75
- Progress in International Reading Literacy Study (PIRLS), 90, 589, 1340, 1452, 1458, 1507nn5–6; data collection and, 1471, 1475; educational policy relation to, 1468, 1490; IRT and, 1451; K–12 education and, 1508n20; literacy and, 1427; population sampling and, 1443; PUF and, 1461–62; test accommodations for, 1449; test administration for, 1478
- progressive matrix, 221
- Project 2025, 25
- Project Essay Grade (PEG), 554, 554f, 555
- projection, prediction and, 793
- Prometric Testing, 516
- promotion exams, 1149–50
- “The Proof and Measurement of Association Between Two Things” (Spearman), 43
- propensity score matching, 605, 801
- Proposed Proactive Training Strategies to Reduce Deleterious Effects on Panelist Judgment and Decision-Making of Judgmental Heuristics and Cognitive Biases (or “Debiasing”), 862t
- Prosser, R., 241
- Protalinski, E., 624
- Provasnik, S., 942, 946, 1481
- proximal consequences, 437n3
- Proyer, R., 1388
- PSAT, 1204
- psychological attributes, 125–26
- Psychological Review* (journal), 42
- psychology: APA and, 52–53, 199, 216, 389–90, 867; Army Air Forces and, 59–60; cognitive, 91; correlational, 37, 42, 43, 63; credentialing in, 1267; decision-making relation to, 834; organizational, 1389; PAD and, 449; psychometrics and, 131; psychophysics and, 46; scaling in, 46, 72
- Psychology Wiki, 840
- psychometrics, 119, 150, 466, 473, 626, 946; in assessment design, 260, 479–80, 481; of Big Five personality model, 1352; CAT and, 584; cheating and, 519; comparability and, 607; computational, 34, 1075, 1320; to conditional inference, 166; construct validity with, 204; CSEM and, 285, 324; data collection and, 1533; data visualization and, 945; measurement error and, 1554–55; PAD and, 449, 450, 471–72, 473; for performance categories, 447; process data and, 614; psychology and, 131; random-item models and, 682; raters and, 525; reflective indicators in, 663; to reliability, 330; RPLDs and, 463, 486; score reporting and, 897, 919, 926–27; SJTs and, 1392; standard setting and, 822, 1316–17; summative assessments and, 489; TBA and, 601, 604; TE and, 984; for test accommodations, 165
- Psychometrika* (journal), 69, 111n17, 112n24
- psychophysical variables, 740
- psychophysics, 34, 36, 37, 46, 1518, 1521–22
- public accountability, 1130–33, 1138
- Public Dissemination of National Assessment of Educational Progress Reading Results in the Metric of Percentage Proficient or Advanced, 1558f
- Public Dissemination of Results From National Assessment of Educational Progress 2022 Reading Assessment, 1556f
- public user file (PUF), 1461–63, 1469
- PUF. *See* public user file
- Puffer, E. S., 126
- Pugh, R. C., 363
- Puhan, G., 766, 778, 795, 1296
- Pullin, D. C., 969
- Purposes of School-Based International Assessments, 1473t
- Putnick, D. L., 1365
- PV methodology, 679–80
- PVs. *See* plausible values
- Pyle, W., 53
- Q-data. *See* questionnaire data
- Qian, H., 614
- Qian, J., 802
- Qian, L., 533
- Q-matrix, 709, 710t
- Q-Matrix Specified for Diagnostic Classification Models, 710t
- quadratic weighted kappa metrics, 556
- quadrature distributions (D method), 346
- qualification exam, for raters, 523
- quality assurance, process data for, 614
- Quality Counts series, 1122
- Qualls, A. L., 245–46, 325, 330–31, 335
- quantitative assessment outcomes, 130–31
- quasi-ipsative data, 1377
- Quellmalz, E. S., 258
- Quenemoen, R. F., 970, 971
- Quesen, S., 240
- questionnaire data (Q-data), 1389
- Quetelet, A., 72

- Quick, N., 970
 Quinn, D. M., 1117, 1167
 Qunbar, S., 131, 555
- Rabe-Hesketh, S., 683, 1535
 Race to the Top (RTTT), 500, 587, 967–68, 1137, 1141–42; growth tracking and, 1151, 1152; SLOs and, 1059, 1060
 racial bias, in facial recognition software, 259
 racial inequity, critical race theory and, 29n2
 racialized groups, educational disparities between, 976–77
 racialized status, 961, 977–78
 racial stratification, wealth relation to, 979–80
 Racine, T. P., 1577n1
 racism, 56–57, 133, 407–8, 968; educational disparities and, 977, 979; systemic, 11, 14, 384
 Radunzel, J., 1206, 1225, 1226, 1231
 Rahman, N., 1232
 Rajaratnam, N., 69, 296, 304
 Raju, N. S., 292–93, 295
 Ramanarayanan, V., 539
 Ramineni, C., 550, 653n4
 Ramsay, J. O., 1359
 Ramsey, P. A., 414
 Randall, J., 11–12, 116, 120–21, 154, 235, 256, 261, 287, 290, 387, 390, 396, 398, 446, 455, 497, 498, 602, 966, 990, 992, 996, 1045, 1168
 Randel, B., 1029
 random effects model, 299
 random-group design, 603
 random-item models, 681–82
 randomly parallel forms, 300, 337
 RAND Reading Study Group, 1544
 Range Performance-Level Descriptor Examples, 462f
 range PLDs (RPLDs), 464t, 465, 466, 859; assessment design and, 471, 476–77, 483, 489; cognition and, 461, 463–64, 480; design patterns and, 467–68; PAD and, 461, 476–78, 480, 492; psychometrics and, 463, 486; test specifications relation to, 474
 Ranger, J., 1380
 ranking, 23, 1376, 1380–81, 1391; ILSAs and, 1485, 1488, 1491; IRT and, 280
 Rasch, G., 77–79, 90, 112n24, 740, 746–47, 1532, 1547, 1561
 Rasch/Guttman scenario scales, 1371, 1373–74
 Rasch ipsative model, 1379
 Rasch model, 36, 78, 93, 345, 1075, 1374; ability estimates from, 746–47; additive conjoint measurement and, 1532–33, 1553; Andrich and, 81; BICAL and, 1534; higher order factor analysis and, 669; interval properties and, 755; IRT *versus*, 79–80, 412–13, 1560; Lexile scale and, 1561; NCLB and, 1074; raters and, 525; scaling and, 742, 1374; vertical scaling and, 789, 791; Woot Math with, 623
 rater-by-task interactions, 68
 rater language background, construct-irrelevant variance and, 237
 rater reliability, 356–58, 365, 1369
 raters, 68, 299–300, 301, 528, 1158, 1368; AI as, 358–59, 542; of behavioral observations, 359–60; bias of, 246, 523, 555, 796; multivariate GT and, 306; precision of, 364; rubrics for, 522–23; scoring inferences and, 246; test scores and, 521, 523; training of, 522, 524–25; true score relation to, 357. *See also* human raters
 Rath, T., 1355, 1356
 rating scale model (RSM), 81, 1360–61, 1366, 1367; bias and, 1391; response style and, 1365; SJTs and, 1384, 1385; validity and, 1362
 Ratwani, R. M., 946
 Raudenbush, S. W., 350, 366, 669, 679, 681, 701
 Rauschenberg, S., 402
 Ravitch, D., 395–96
 Rawls, J., 196, 234
 raw scores, 283, 289, 310, 340, 754, 927; CSEM and, 331–32, 335; response time and, 615; TBA and, 600, 601; variance and, 320
 raw-to-scale score conversions, 336–37, 347, 752–53, 785–86
 Ray, M. E., 1295
 Raykov, T., 349
 Raymond, M. E., 1144–45
 Raymond, M. R., 852, 854, 1132–33, 1266–67, 1268, 1288, 1289–90, 1291, 1313, 1314
 RD. *See* receptive dichotomous items
 reading comprehension, 25, 118, 1542f, 1556f, 1558f, 1565t; design patterns for, 467; log files and, 612; NAGB relation to, 1544, 1565, 1566; PAD and, 497; sociocultural theory of, 1545f; test scores for, 1578n9; TISM framework and, 1516, 1541, 1542, 1543–44, 1545–49, 1555
 Reagan, R., 62
 real-effort tasks, 1387
 real-time learning interactions, intelligent tutoring systems and, 632
 Reardon, S. F., 136, 802, 978, 1132, 1148, 1149
 reasoning, 129, 447
 Reback, R., 1147
 Rebell, M. A., 402–3, 408, 1145–46
 receptive dichotomous items (RD), 719–20, 721–23
 receptive modality, 718–19
 Reckase, M., 78, 822, 850–51, 861, 877
 Reddy, V., 1234
 Reddy, Y. M., 1031
 reduction operator matrices, 694
 Reese, L. M., 1220
 Reese, T. J., 84, 946
 Reeves, R. V., 1345, 1350–51
 reflective indicators, 663
 regressed score estimate, 343–44
 regression, 39–40, 58, 536, 548, 685, 1154; differential prediction via, 438n11; prediction and, 785; selection decisions and, 421; of true score, 764
 regression analysis, 388, 423, 659
 regression coefficient, 314, 716
 regression invariance, 416–17
 regression modeling, 711
 regulation, 1266, 1270–73
 Rehabilitation Act (1973), 142, 1280
 Reid, A., 7
 Reid, J. B., 1313, 1314
 Reid, J. R., 852, 854
 Reinsch, C., 543
 Reise, S. P., 666, 704, 742, 744, 746, 1361, 1421n4
 Relationship Between Large-Scale Assessment and Society, 128f
 relative error variance, 302–3
 relevance, 121, 979, 1438, 1481
 reliability, 14, 63–66, 69, 285, 301, 351; administration conditions and, 364; of admissions testing, 21–22, 1223, 1223t, 1224; of AI, 551; assessment development and, 767; AVs and, 1372; BARS and, 1375; of behavioral observations, 359–60; for beta-binomial model, 343; of Big Five personality model, 1352, 1356–57; of coefficient alpha, 294; confidence intervals and, 349, 357, 360; COVID-19 pandemic and, 795; in credentialing, 23, 1298–1300; CSEM and, 291, 362, 367; CTT and, 64–66, 278, 280–81, 286, 412, 1359–60, 1361; data collection for, 290; DCM and, 348; equating relation to, 768, 796; error variance and, 363; estimation and, 279–80; expected error variance relation to, 283; for forced-choice methods, 1380; of grading, 1041, 1043–44; GSB and, 293; in GT, 66–67, 92, 214, 278, 280–81, 297–98, 361; human-scored responses for, 528; of independence assumption, 674; of interim assessment, 1033; interrater, 492; in IRT, 280–81, 309–10, 318; level-specific, 355; in licensure, 1269; normal approximation procedure and, 341–42; number-correct scores and, 352, 363; observed score and, 281, 1061; parallel-forms coefficient and, 316; precision and, 366; psychometrics to, 330; rater, 356–58, 365, 1369; replication and, 15, 279; of score interpretation, 897; score weights and, 743, 744; SJTs and, 1385; Spearman-Brown formula and, 45–46; of SpeechRater, 556; of speeded assessments, 353–54; of subscores, 920; of test scores, 200, 279, 1297, 1301; validity and, 194, 195, 198; in value-added models, 350; variance and, 69

- The Reliability and Validity of Tests* (Thurstone), 65
- reliability coefficients, 45, 279, 317–18, 354–55, 1526; absolute error variance and, 342–43; for adaptive testing, 352, 366–67; AI and, 358; coefficient alpha and, 297, 349, 361; conditional SEM and, 320; consistency of, 360; CSEM and, 324–25; CTT and, 285, 288, 302–3, 311; general-purpose, 296; GT and, 364; IRT and, 288, 348, 352–53, 360; for linear composite, 321–22; in Livingston and Lewis procedure, 343; MST and, 352; SEM and, 287–88, 289–90, 1554; Spearman-Brown formula and, 292; structural equation models and, 349; tau-equivalent forms and, 294–95; in technical manuals, 362; test information function and, 316; test-retest, 290–91, 364–65; variance-ratio definition and, 289, 315
- reliability generalization (RG), 355–56
- Reliability of Undergraduate and Graduate Admissions Tests, 1223*t*
- reliability theory, 63, 69
- Remaining Steps in Assessment Design, 465*f*
- remediation, credit-bearing courses *versus*, 1233–34
- Remmers, H. H., 45
- remote learning, 136
- remotely administered exams, 6, 15, 384, 400–401, 402
- remote proctoring, 402, 516, 589, 634, 1191
- REMSD. *See* root expected-mean-square difference
- replication, reliability and, 15, 279
- report cards, 1130–31, 1132–33, 1134, 1340
- report development, for score reporting, 931–32
- representational measurement theory, 1532
- research agenda, for standard setting, 874–76
- Resnick, D. P., 1047, 1121
- Resnick, L. B., 1047, 1121
- resolved score, human raters and, 557
- respectful treatment, fairness and, 400
- response formats, 612, 1455, 1477–79, 1510; Army Alpha and, 53; in large-scale testing, 130; PISA, 1448; ranking, 1376
- response probability (RP), 683, 685, 759, 868–69
- response process, 241, 254–55, 613, 653*n*2, 1377; TBA and, 611–12, 635; test scores and, 216
- response rates, 1444–45, 1462, 1490–91
- response style, RSM and, 1365
- response time, 354, 519, 583, 614, 749, 1388; data collection for, 1307, 1308; IRT and, 80, 1379–80; raw score and, 615
- Response-to-Intervention program, 1053
- results report. *See* score reporting
- Revelle, W., 361
- Revicki, D. A., 1361
- Revilla, M. A., 1358–59
- Reynolds, K., 18
- RG. *See* reliability generalization
- Rhoades, K., 1048
- Rice, J. M., 85–86, 1522, 1523–24
- Richardson, Marion W., 34, 36, 65, 66, 92, 294, 1342, 1343, 1362, 1526
- Richardson extrapolation method, 703
- Richer, A. D., 1198
- Ricker-Pedley, K. L., 521
- Rico, R., 875
- Riconscente, M., 456
- Rieger, S., 1346
- Riggan, M., 1036
- Rigol, G., 1199, 1200, 1263*n*3
- Rijmen, F., 632, 783
- Rimfeld, K., 1346
- Rios, J. A., 974, 1344
- Rippetoe, R., 1194
- Rissanen, J., 698
- Ritter, S., 623, 1070, 1074
- Ritzen, J., 1425, 1468, 1470, 1484
- Rivas, A., 616
- Rivera, C., 610, 973
- Riverside, 86
- Rivkin, S. G., 1141
- RM. *See* Robbins-Monro recursions
- RMSD. *See* root-mean-square difference
- RMSEA. *See* root-mean-square error of approximation
- Roach, A. T., 968, 994
- A Road Map for Audience-Based Research, 903*f*
- Robbins, H., 690
- Robbins, S. B., 1342, 1362
- Robbins-Monro recursions (RM), 690–91
- Roberts, B. W., 1342, 1343, 1346
- Roberts, G. O., 691, 698
- Roberts, J., 1200
- Roberts, J. S., 1377, 1378
- Roberts, M., 943
- Roberts, S. O., 116
- Robertson, S., 547
- Robin, F., 584, 585, 586–87
- Robins, R. W., 1342, 1343, 1362
- Robinson, D. H., 57, 81–82
- Robinson, J. P., 1474
- Robinson, V., 1141
- Roch, S. G., 524
- Rock, D. A., 1389
- Rockoff, J. E., 1141, 1144, 1388
- Roda, A., 1135
- Roderick, M., 1137, 1141, 1149
- Rodriguez, M. C., 19, 135, 235, 387, 398, 494, 511*n*2, 515, 516, 517, 597, 970, 979, 983, 985, 992, 994, 1018, 1164, 1166, 1168, 1230, 1280, 1283–84, 1357, 1449
- Rodriguez, O., 1233
- Roduta Roberts, M., 931, 937
- Roeber, E., 1159
- Rogers, C. M., 970
- Rogers, H. J., 411
- Rogoff, B., 116, 126, 152, 396
- Rogosa, D., 350, 360
- Romboy, D., 1272
- Romme, G., 451
- Roohr, K. C., 1239, 1241
- root expected-mean-square difference (REMSD), 420
- root-mean-square difference (RMSD), 419–20
- root-mean-square error of approximation (RMSEA), 678, 697, 699, 707, 710
- Rosdahl, A., 1474
- Rose, D. H., 124, 244, 966, 967, 1166
- Rosén, M., 1471
- Rosen, Y., 259, 260
- Rosenbaum, P. R., 605
- Rosenberg, A., 1518, 1577*n*2
- Rosenfeld, M., 1290–91
- Rosenshine, B. V., 1071
- Rosenthal, J. S., 691
- Rosenthal, L., 1140
- Rosenthal, R., 1025
- Ross, J. A., 1031
- Ross, L., 846
- Rosyd, H. A., 520
- Roth, W.-M., 119, 121, 126, 131, 149, 161–62
- Rothman, R., 1047, 1141, 1157
- Rothmann, S., 1342
- Rothstein, J. M., 407
- Rothstein, R., 140
- Rottenstreich, Y., 838
- Rotter, J. B., 1355
- Rouse, C. E., 1145, 1341
- Rowe, K. J., 1068
- Rowley, K. J., 1487
- Royal, C., 1145
- Royal Geographical Society, 38
- Royer, J. M., 968
- RP. *See* response probability
- RPLDs. *See* range PLDs
- RSM. *See* rating scale model
- RStudio, 945
- RTTT. *See* Race to the Top
- Rubel, L., 978
- Rubin, D. B., 605, 656, 679, 765, 1453
- Rubin's multiple imputation theory, 679
- Rubinstein, Y., 1341
- rubrics, 534, 556, 1031, 1239; construct definition in, 620, 621; for raters, 522–23; test scores and, 550–51
- rubric validation, 528
- Ruch, G. M., 45, 86
- Rudner, L. M., 345–46, 1298
- Ruiz-Primo, M. A., 118, 119, 131, 154, 485, 1023, 1028, 1030, 1070
- Rulon, P. J., 198, 200, 292
- Rupp, A. A., 162, 442, 627, 668
- Rupp, N. G., 1390
- Russell, M., 11, 116, 218, 590, 592, 596, 601, 961, 962, 984, 986, 988, 996

- Rust, K., 1436, 1443, 1457
 Rutkowski, D., 1472, 1474, 1478, 1479, 1486
 Rutkowski, L., 1446, 1456, 1467, 1472–73, 1478, 1479, 1483
 Rutschow, E. Z., 1236
 Ryan, J. E., 1151
 Ryan, J. M., 901, 908, 931, 935, 937, 938, 979, 1126
 Ryan, K., 217, 219, 220
 Rychen, D. S., 1434
- Sackett, P. R., 364, 407, 1226
 Sadler, D. R., 1022, 1030–31
 Safe Harbor, 1124
 Saka, N., 518
 Salgado, J. F., 1342, 1343, 1377, 1378, 1381
 Salganik, L. H., 1434
 Salmani Nodoushan, M. A., 986
 Salzburg Global Seminar, 1340
 Samejima, F., 80, 313, 315, 682, 1378
 Samelson, F., 53
 Sample Graph of Curriculum-Based Measurement (CBM) Measures Collected Over 27 Weeks of Instruction, With an Instructional Change Occurring After Week 1054f
 Sample Report Section With Lexile-Based Book Suggestions, 900f
 Sample Theory of Action (TOA), 484f
 Sanchez, C. E., 1042
 Sanchez, E. I., 1195, 1197, 1231
 Sánchez-Meca, J., 355
 Sandene, B., 587, 601
 Sanders, R. E., 832
 Sands, A., 1510n62
 Sands, W. A., 61, 585
 Sanger, Margaret, 56
 Saretzky, G. D., 385
 Sarikaya, A., 924
 Sartain, L., 1141, 1143
 Sass, R., 1382
 SAT, 36, 85–88, 112nn29–30, 898, 1207t, 1264n10; academic performance relation to, 1226–27; accessibility and, 401; for accountability testing, 231; ACT relation to, 1205–6; admissions testing and, 1199; alignment and, 1159; Army Alpha relation to, 92; battery scaling and, 757, 792–93; concordance and, 788–89; COVID-19 pandemic and, 1191, 1212, 1213; decision-making and, 133–34; equating and, 765, 766, 785–86; eugenics movement relation to, 385; flagging policies for, 405; HSEEs and, 1139, 1140; HSGPA relation to, 1198; linking invariance and, 419, 420; machine scoring of, 1202; *A Nation at Risk* on, 1120; PBA and, 600; as placement test, 1234; prediction invariance and, 418; regression analysis for, 423; remotely administered exams, 6, 13; scale scores for, 750, 752, 753, 754; SEs and, 781; social justice relation to, 1192; TBA and, 586, 607; test accommodations for, 968; test-optional policies relation to, 1196; test scores on, 1203–4, 1211; UC and, 406–7, 1193, 1194; validity of, 1231; vertical scaling and, 789
 Satorra, A., 349
 Saucier, G., 1353
 Saul, S., 1195
 Savalei, V., 1421n3
 Sawyer, R. L., 338, 610, 1225, 1226
 SBTs. *See* scenario-based tasks
 scale-aligning methods, 762, 768–69, 777, 779, 793; concordances, 786, 788–89, 1205; equating and, 770, 775, 783–84; linear, 765, 784–85
 scale anchoring, 72, 759, 920, 1371
 scale scores, 18, 284, 322, 753, 1437, 1547; for ACT, 750, 752, 754; CSEM and, 289, 330, 331, 332, 333, 335, 336–37; for GRE, 750, 754, 1219; IRT and, 335, 367, 759, 764; in K–12 assessment, 755, 760; linear transformation and, 319; norms and, 757–58; for placement tests, 361; score interpretation of, 1556, 1557, 1558–60; score reporting and, 918–19, 927; test performance and, 738, 750; TISM framework and, 1558, 1560, 1561; truncation and, 751
 scale transformations, 762–63
 scaling, 73–74, 1523, 1547, 1561; background questionnaire and, 1451, 1455, 1459, 1478–79; battery, 756–57, 792–93; cognitive skills and, 1479; data collection for, 738t; equating relation to, 16; Hieronymous, 791; of interim assessment, 1033; IRT and, 71, 74–75, 93, 740, 743, 745, 1441–42, 1489; for large-scale testing, 737, 750, 760, 802; in measurement theory, 741; personality tests and, 1357; in psychology, 42, 72; Rasch model and, 742, 1374; for reading comprehension, 1547; score reporting and, 736, 737, 738–39, 804; self-report ratings and, 1358–59; TISM framework and, 1538–39; vertical, 755, 789–90, 791–92
 Scalise, K., 157, 927–28, 1065, 1072, 1073, 1076
 SCANS Commission report, 1344–45
 scenario-based tasks (SBTs), 17, 594, 595f, 595–96, 598, 610
 Schaeffer, G. A., 601
 Schaeffer, K., 979–80
 Schaepers, P. C., 1386
 Schalet, B. D., 793
 Schanzenbach, D. W., 1148, 1340
 Schaubroeck, J., 1369
 Schell, K. L., 1377
 Schemo, D. J., 1203
 Scherer, R., 1471
 Schifter, L. A., 980
 Schildkamp, K., 1045
 Schläpfer, F., 1387
 Schleicher, A., 1485
 Schleifer, T., 27
 Schmeiser, C. B., 470, 472, 473
 Schmidt, F. L., 393, 515, 1389
 Schmitt, A. P., 414
 Schmitt, K., 1270, 1271
 Schmitt, M., 1388
 Schmitt, N., 1365, 1382, 1383, 1385, 1386
 Schmitt, T. A., 353
 Schneider, M. C., 16, 460, 463, 467, 477, 485, 1029–30, 1061, 1070, 1134–35
 Schnellert, L., 1022
 Schnepf, S., 1480
 Schnipke, D., 855
 Schoenfeld, A., 1031
 Schoenig, R. W., 518, 519
 Schofield, A., 547
 Scholastic Aptitude Test, 87, 112nn29, 385, 1202, 1203, 1222. *See also* SAT
 Scholastic Aptitude Test for Medical Students, 1222
 Scholastic Assessment Test, 112nn29, 1203
 School-Based International Tests of Educational Achievement: Scope and Timing, 1428t
 school choice, 1122–23, 1133–34, 1135
 school performance, 393, 1148
 Schratz, M., 158
 Schudson, M. S., 1202
 Schultz, M. T., 559
 Schulz, E. M., 345
 Schum, D. A., 450
 Schunk, D. H., 1355
 Schuurman, N. K., 350
 Schuwirth, L. W. T., 1321
 Schwager, I. T., 1229
 Schwartz, G., 678, 1065
 Schwartz, S., 8
 Schwendimann, B., 924
 Schwerdt, G., 1149
 Schwippert, K., 1468
 science, technology, engineering, and mathematics (STEM), 1196, 1205–6
The Sciences of the Artificial (Simon, H. A.), 451
Scientific American (journal), 7
Scientific Management in Education (Rice), 85
 scientific realism, 1518
 Sconing, J., 1206
 Scopes, John T., 7
 Scopes trial, 7
 score comparability, 124, 235, 255, 601–2, 606
 score gaps, in large-scale testing, 118
 score inferences, 213, 245, 447, 514, 583
 score interpretation, 14, 67, 253, 278, 929–30, 1514; consequences of, 232, 233; construct-irrelevant variance relation to, 517, 562; evidentiary argument for, 485; fairness of, 195, 196, 387; in PAD, 451; reliability of, 897; of scale scores, 1556, 1557,

- 1558–60; in score reporting, 916–17;
sociocultural context for, 118; TBA and,
254; TE and, 984; test use and, 208–11,
217, 239; validity of, 125, 192–94, 197,
207, 210, 212, 217, 218, 223, 224, 239–40,
960, 961, 1280, 1292, 1305, 1308, 1553
score-level comparability, 602–3
score reporting, 23, 62, 908, 909–11, 922,
1295; assessment development relation
to, 898, 946; as communication, 901–2,
917; contextual information for, 921;
dashboard for, 923–25, 928; as data story,
911–12; data visualization for, 945–46;
equating and, 736, 760, 804; field testing
for, 932–36; for K–12 assessment, 899,
899f, 900, 930, 943; machine learning and,
944; NAEP and, 902, 916, 927, 1119–20,
1163; narrative patterns for, 911–12, 913,
913f, 914; for next-generation assessments,
946; performance standards and, 736–37;
process data for, 942–43; psychometrics
of, 897, 919, 926–27; scale scores in, 918–
19, 927; scaling and, 736, 737, 738–39,
804; score interpretation in, 916–17;
stakeholders and, 896, 902, 931; of sub-
scores, 900, 919–20; technology-based,
943–45; after test administration, 925;
test performance and, 915–16; test pur-
poses relation to, 926; validity and, 896,
897, 903, 906, 907–9, 937, 947
score weights, 743–44
scoring algorithms, CLR in, 256
scoring inferences, 211, 246, 249
Scott, E. L., 663
Scott-Clayton, J., 1234
Scoular, C., 259
Searle, S. R., 681
SEC. *See* Surveys of Enacted Curriculum
second-language learners, 122–23
Secretary's Commission on Achieving Nec-
essary Skills, 1434
Section 504, of Rehabilitation Act, 142
SEE. *See* standard error of estimation
SEEDs. *See* standard errors of equating
differences
Segal, C., 1342, 1343, 1370, 1387, 1388
Segall, D. O., 352, 798
segregation, 397, 977–79
Seixas, P., 459
SEL. *See* social-emotional learning
SEL Assessment Guide, 1366
Selected Milestones in the History of
Standard Setting, 825t
Selected Screens From the NAEP TEL 2014
Scenario-Based Task *Bike Lane*, 595f
selection decisions, 233–34, 420–22
selection procedures, 59, 85
selection tests, 16, 437, 472, 1274, 1276;
equity and, 236; test scores for, 233–34
self-assessment, 1031, 1067
self-report ratings, 1358–59, 1366, 1391
self-selection bias, 553
Seligman, M. E. P., 1356
Sellar, S., 1484
SEM. *See* standard error of measurement
Semega, J., 980
sensitivity guidelines, 395–96, 498, 990–91
sensory intensity, 71–72
Sentence Map, 1373–74
Seo, D. G., 352
serious-based games, 530, 532
SEs. *See* standard errors
SES. *See* socioeconomic status
Setati, M., 135
Settle, R. B., 1358
Settles, B., 624
Seybert, J., 1380
Sgammato, A., 849
SGP. *See* student growth percentile model
Shaffer, D. W., 632
Shafik, M., 10
Shafir, E., 127
Shaftel, J., 985
Shanedling, J., 984
Shanks, D. R., 841
Shapiro, D., 1241
Shapiro, I., 9
Sharkey, P., 979
Sharon, A. T., 404
Shavelson, R. J., 35, 66, 134, 163, 298, 598,
681–82, 1067, 1528
Shaw, E. J., 1226, 1549, 1550
Shaw, R., 855
Shaw, S. D., 194, 206, 212, 222, 224
Shaw, S. R., 127
Shea, J. A., 449, 1314
Shealy, R. T., 413
Sheehan, K. M., 1555
Shelton, P. D., 1209
Shepard, L. A., 14, 136, 192, 193, 196, 206,
208, 219, 220, 221, 225, 237, 239, 252,
389, 391–92, 404, 408, 425, 442, 444, 485,
632, 848–49, 982, 986, 988, 1029, 1033,
1034, 1035, 1047, 1056, 1071, 1075, 1118,
1131–32, 1168, 1316–17
Shermis, M. D., 15, 249, 358, 532, 533, 534,
538, 539, 553, 555, 556, 558, 583, 616, 618,
658, 961
Sherry, D., 1537, 1553
Shi, D., 678
Shields, J., 59
Shimamoto, H., 552
Shimberg, B., 1271
Shin, J., 547, 1055
Shindoll, R. R., 1312
Shivakumar, N., 521
SHL Group, 1376
Shock, N. W., 45
short-answer scoring, 538
Shriner, J. G., 1056
Shrout, P. E., 349
Shulman, L., 1071
Shultz, M. M., 1214, 1355, 1374, 1375
Shultz, P. K., 1169
Shute, V. J., 493, 623, 626, 1025
SIBTEST procedure, 413
Siemens, G., 928
Sievert, C., 945
signal detection theory, 839
sign-trait interpretations, 449
Sijtsma, K., 162, 668, 1532, 1547, 1551–52,
1553
Simmons, C., 916
Simon, H. A., 131, 451, 741, 759, 841
Simon, J. P., 534
Simon, M., 135
Simon, M. A., 1062
Simon, T., 36, 48–50, 51, 93, 110n8, 132, 739,
1520
Simone, S., 1233, 1234
simple-structure MIRT (SS-MIRT), 320–21,
345
Simplified Sample of a K–12 Results Report,
899f
Simplified Sample Report With Annotations
Highlighting Narrative Patterns, 914f
Sims, D., 1155
simulation-based task, 157, 634
simulation-enabled performance tasks,
596–97, 598, 599, 610, 612
Singer, J. D., 140, 1152, 1425, 1473, 1474,
1485, 1487, 1510nn59–60
Singh, N., 526
single-group designs, 603
singular value decomposition, 544f
Sinharay, S., 348, 697, 759, 776, 910, 919, 920,
1296, 1433, 1436, 1450, 1452, 1457, 1458,
1508n32
Sinnema, C., 1141
Sireci, S. G., 121, 145, 147, 154, 199, 240, 387,
390, 516, 517, 591, 602, 609, 610, 803, 868,
905, 915, 961, 970, 984, 998, 1156, 1157,
1158, 1159, 1308
Sirotkin, 547
Sisk, V. F., 1347–48
Sitgreaves, R., 112n24
situated learning theory, 194, 244
situational judgment tests (SJTs), 23,
1382–86, 1389, 1392
Skaggs, G. E., 844, 860, 867, 868, 869, 873,
894n9
Skinner, B. F., 622, 1118, 1543
Sklar, J., 418
Skorupski, W., 828, 843, 851, 866
Skrondal, A., 683, 1535
Slaney, K., 1549, 1550, 1577n1
Slater, S. C., 910, 916, 929, 931, 932, 937, 944
Slavich, G. M., 1390
sLDA. *See* supervised LDA
Sleeman, D., 623
SLO. *See* student learning outcomes
Sloane, K., 1064, 1065, 1069, 1070, 1073
Slomp, D., 11–12

- SLOs. *See* student learning objectives
- Slovic, P., 846
- Smarter Balanced Assessment Consortium, 556–57, 587, 598, 599–600, 601, 608, 609, 654n5, 855, 1051, 1136, 1158, 1159, 1193, 1541, 1542, 1543f, 1546, 1548–49, 1554, 1561
- SMEs. *See* subject matter experts
- Smiley, W., 1296
- Smith, A. Z., 1044
- Smith, E., 1118, 1121
- Smith, I. L., 1311
- Smith, J., 1198, 1209
- Smith, J. K., 1025, 1046, 1300, 1316
- Smith, K., 406–7
- Smith, M., 1510n59
- Smith, M. S., 1156
- Smith, R., 789
- Smith, R. A., 935, 1027, 1053, 1055–56, 1059
- Smith, R. L., 1316
- Smith, S. S., 1144
- Smith, W., 1425, 1484
- Smith v. Regents of University of California* (2019), 401, 406–8, 1194
- Snapshot of Design Pattern With Manipulable Features of Complexity (Difficulty Drivers), 469f
- Snapshot of Integrated National Assessment of Educational Progress Technology and Engineering Literacy Specifications, 476f
- Snow, C. E., 561
- Snow, R. E., 449, 457, 500
- Soares, J. A., 1209
- social-desirability bias, 1362
- social-emotional learning (SEL), 8, 1340–41, 1348, 1350, 1391; in K–12 education, 1074; personality traits and, 1346
- social inequities, 137, 138, 140, 408, 426
- social justice, 4, 7–13, 20, 196, 197, 220; admissions testing and, 22, 1192; COVID-19 pandemic relation to, 1192; culturally responsive assessment relation to, 235–36, 237, 246; diversity relation to, 245; selection decisions and, 422; sociocultural context and, 116; sociocultural learning theories for, 252
- societal needs, culture relation to, 129
- Société libre pour l'étude psychologique de l'enfant, 48
- Society for Human Resource Management, 1345
- Society for Industrial and Organizational Psychology, 1282
- sociocognitive learning theories, 239–40, 242–43, 1168
- Sociocultural Considerations in the Assessment Triangle, 160f
- sociocultural context, 117, 134, 148, 237–39, 986; assessment design and, 120, 125, 166–68, 444; in assessment interpretation, 118–20, 121, 130; assessment performance relation to, 123, 139, 156–57, 165; for decision-making, 151, 152; fairness and, 120, 125, 136, 161–62, 239, 988; international assessments relation to, 138; for interpretation model, 158, 159; language in, 135; in large-scale testing, 118, 128, 128f, 129; for learning outcomes, 127, 148; MC relation to, 165; for observation model, 154, 155; social justice and, 116; standardization relation to, 124; for standard-setting method, 18–19; teaching expectations relation to, 141; validity and, 119, 125
- sociocultural learning theories, 252, 261, 982–83
- sociocultural theory, 553–54, 1545f
- socioeconomic status (SES), 230–31, 393, 425, 454–55, 1343, 1471; achievement gaps relation to, 231, 977–78; admissions testing and, 393, 407–8; as construct-irrelevant variance, 980; educational disparities relation to, 979; ILSAs relation to, 1479–80; of panelists, 870; in PUF, 1462; UC and, 1192
- sociopolitical components, 132
- Soderstrom, N. C., 1025
- soft-max, 617, 653n1
- software-based data collection, 724
- Sohn, H., 1147, 1148
- Sokal, M. M., 51
- Soland, J., 1345, 1480
- Solano-Flores, G., 13, 116, 117, 119, 120, 121, 122, 125, 131, 135, 138, 143, 145–47, 149, 154, 156, 163, 165, 235, 236–37, 240, 241, 243, 455, 497, 498, 611, 660, 724, 988, 992, 999, 1018, 1274, 1281
- Solberg, L. I., 1049
- SOLO taxonomy, 459
- Sörbom, D., 295
- Sorbonne entrance examination, 385
- Sotaridona, 519
- Soto, C. J., 1389
- Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers English Language Arts/Literacy, 470f
- Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers Math, 471f
- South Carolina Department of Education, 444
- Sowell, Thomas, 9
- Sparks, J. R., 1239
- Spearman, C., 34, 35–36, 37, 40, 42, 43–46, 50, 55, 63, 92, 93, 291, 672, 1520, 1522, 1524, 1526, 1530
- Spearman-Brown formula, 44–45, 291, 292, 293, 1360, 1526
- special education, 146, 147, 388
- special needs, 144–45, 146–48. *See also* disabilities; English learners; students with disabilities
- special populations, 723–94, 1163–66, 1230
- specific objectivity, 78, 1532
- SpeechRater, 540–42, 556
- speech recognition technology, 526
- speech-to-text translator programs, 539
- speeded assessments, reliability of, 353–54
- Speekenbrink, M., 841
- Spellings, M., 1151, 1190–91
- Spillane, J. P., 1146
- Springer, 1311
- Sputnik*, 61
- squared-correlation definition, 288
- squared-error loss methods, 338
- Srivastava, S., 1342, 1350, 1352, 1356, 1361
- SSAT-II. *See* State Student Assessment Test, Part II
- SS-MIRT. *See* simple-structure MIRT
- stability measures, 202
- Stack, L., 26
- Stafura, J., 1544
- Staiger, D. O., 1154, 1155
- stakeholders, 1268, 1299, 1313, 1315, 1438, 1439; background questionnaire relation to, 1441–42; credentialing and, 1277–78, 1314; decision-making and, 964; for educational policy, 1433, 1435, 1436; grading relation to, 1042; ILSAs relation to, 1426, 1471, 1485, 1489; score reporting and, 896, 902, 931
- standard error of estimation (SEE), 310
- standard error of measurement (SEM), 284–85, 366, 754, 757, 1299; in CTT, 289–90, 310, 324; IRT and, 675, 1554; reliability coefficients and, 287–88, 289–90, 1554. *See also* conditional SEM
- standard errors (SEs), 345, 703, 711–12, 781–82
- standard errors of equating differences (SEEDs), 782
- standardization, 412, 413, 426, 516, 609, 977; administration conditions and, 242, 249, 607; culture relation to, 236–37; fairness and, 124, 135, 161, 164–65, 1303; MH, 411–13, 987; NCLB and, 1164; selection decisions and, 421; of summative assessments, 658; TBA and, 589; validity relation to, 124–25, 258, 514–15
- standardized LD, 706
- standardized root-mean-square residual, 678
- Standardized Testing Act, 404
- Standardized Testing Task Force, 407, 1192–93, 1194, 1198, 1210
- standardized tests, 62, 134, 1143, 1222, 1384; fairness of, 384, 392, 1230; for intelligence testing, 386; at UC, 1194. *See also* specific tests

- Standard Process Approach for Automated Speech Scoring, 539f
- standards-based grading, 1042, 1044
- standards-based reform theory, 1156
- standard setting, 460, 855, 870, 1315; bias in, 846, 1319; cognitive and social model of, 828, 830f, 847–48, 873–74, 879; cognitive-judgmental task for, 823–24, 829, 834, 835f, 837; cognitive process in, 841; for credentialing, 824, 833, 853, 859, 879, 1270, 1309–12, 1317; cut score and, 756; embedded, 722, 877–78; expert panels for, 852–53; group discussions in, 871; human judgment for, 823, 828–29, 834; impact data for, 850–51; IRT relation to, 878, 879; for licensure, 1311; panelists for, 1313–15; policy decisions and, 1316, 1317–18; as policymaking, 820n1, 822, 851; probability judgments for, 837–38; psychometrics and, 822, 1316–17; research agenda for, 874–76; RPLDs and, 463; Selected Milestones in the History of, 825t; target PLDs and, 465; test content and, 756, 849, 992; threshold-signal detection judgments for, 839–40; value judgments for, 840, 1318–19
- standard-setting design, 848, 853–54, 871–72, 876, 877; cognitive-judgmental task relation to, 856, 863; panelists relation to, 850, 852
- standard-setting facilitator, 848
- standard-setting method, 18–19, 23, 133, 224–25, 859
- Standards for Educational and Psychological Testing* (AERA), 151, 192–93, 195, 197, 203, 204, 211, 212–13, 216, 330, 337, 352, 361, 362, 386, 387, 388, 389–92, 394, 395, 398, 400, 405, 406, 419, 426, 437n4, 443, 514, 550, 583, 601–2, 853, 904, 904f, 907, 908, 910, 921, 925–26, 960, 964, 968–69, 973, 990, 997, 1020, 1116, 1150, 1233, 1274, 1279, 1282, 1287, 1288, 1298, 1310, 1316, 1318, 1388, 1549, 1550
- Standards for the Accreditation of Certification Programs, National Commission for Certifying Agencies, 1300
- Stanford Achievement Test, 86
- Stanford-Binet scale, 52, 132, 134
- Stanford Named Entity Recognizer, 557
- Stankov, L., 1354, 1388, 1421n1
- Stanley, J. C., 85
- Stapleton, L. M., 872
- Star, S. L., 122
- Starch, D., 1043
- Stark, S., 749, 1377, 1378–79
- Starling, M., 1031
- state-based linking, 802
- stated-effort tasks, 1387
- state-mandated accountability tests, TBA for, 586–87
- state-specific learning standards, 471
- State Student Assessment Test, Part II (SSAT-II), 397
- state summative assessments, 20, 1054, 1073, 1136–37; as assessment for teaching and learning, 1047, 1051; data collection and, 1048–49
- state testing, 219–20, 231–32, 252; alignment and, 493, 1157–58, 1159; undergraduate admissions relation to, 1210
- Statistical Theories of Mental Test Scores* (Birnbau), 77
- statistical theory, psychology and, 46
- stealth assessment, 493, 494
- Stecher, B., 1345
- Stecker, P. M., 1053, 1055, 1056, 1057–58, 1059
- Steedle, J., 463, 1226
- Steele, C. M., 138, 154
- Steele, S., 9
- Stein, B., 520, 521
- Stein, Z., 1564
- Steinberg, L., 81, 83, 112n26, 660, 667, 704, 1360, 1532
- Steinberg, M. P., 1142, 1143
- STEM. *See* science, technology, engineering, and mathematics
- Stemler, S. E., 1355
- Stenner, A. J., 1555, 1560
- Stern, J., 785
- Stern, William, 111n10
- Sternberg, R. J., 1383, 1385
- Stevens, S. S., 71, 740–41, 742, 755, 1515–16, 1521–22, 1552
- Stiggins, R. J., 534, 1020, 1043, 1071
- Stigler, S., 38, 46, 67, 68
- Stilwell, L. A., 1229
- Stocking, Martha, 82, 585
- Stocking-Lord equating method, 82
- Stoeffler, K., 1072
- Stone, C. A., 220, 221, 228, 229, 240, 257
- Stone, E., 240, 254, 969, 970, 984
- Stone, T. H., 1375–76
- Stonehill, R. M., 62
- Stonewall Inn, 10
- Stouffer, S. A., 60
- Stout, W. F., 413
- Strain-Seymour, E., 609, 984
- stratified coefficient alpha, 297
- stratum-level scores, 297
- StrengthsFinder 2.0, 1356
- Stricker, L. J., 423, 987, 1389
- Strietholt, R., 1471
- Strivers method, 423
- Stronge, J. H., 1141
- Stroud, J. B., 518
- structural equation models, 15, 349, 678, 681
- Strunk, K., 1145
- Struthers, K. S., 1145
- Stuart, A., 322, 357
- student achievement, 1142, 1143–46, 1148–50
- student growth percentile model (SGP), 1152, 1153–54, 1155
- student learning objectives (SLOs), 20, 1052, 1059, 1060, 1061–62
- student learning outcomes (SLO), 1037, 1237–38, 1239, 1243; accountability and, 1190–91, 1240; formative assessments relation to, 1031
- student navigation experience, 498–99
- Students for Fair Admissions v. President and Fellows of Harvard College* (2023), 1200–1201
- students with disabilities (SWDs), 62, 407, 454–55, 970, 980–81, 999; accessibility for, 387, 444, 967–68, 971, 975; ADA and, 135, 139, 142; alternate assessments for, 511n2, 832, 972; CBM for, 1053, 1058; consequences for, 228–29; construct-irrelevant variance and, 166; fairness and, 123, 388; flagging policies and, 405–6; formative assessments and, 1022; ILSAs and, 1448; innovative item types and, 597; in K–12 education, 968; NCLB and, 63, 143–44; poverty and, 980; standardization and, 124; test accommodations for, 141, 145–47, 165, 167, 254, 399, 401–2, 516, 969–70; test-based accountability for, 1163–64, 1165–66; UDL and, 244
- Studley, R. E., 407, 423
- A Study of American Intelligence* (Brigham), 56, 111n14, 386
- Study on Social and Emotional Skills, OECD, 1356
- Su, W., 422
- Suarez-Alvarez, J., 1351
- subject matter experts (SMEs), 1292
- Subkoviak, M. J., 341, 342, 343–44, 1298
- subpopulation invariance evaluations, 780, 782–83
- subscores, 20, 900, 919–20, 1295–96
- substitution heuristic, 860
- Sulsky, L. M., 524
- Summary of Alt ELPA Grade 2–3 Item Characteristics, 719t
- Summary of Data Collection Designs for Scaling, Equating, and Linking, 738t
- Summary of Nonequating Types of Linking, 787t
- Summary of Recent Linking Examples, 787t
- Summary of Types of Equating, 761t
- summative assessments, 442, 443–44, 658, 1051, 1130, 1307; formative assessments compared to, 1019, 1076; instructional decisions relation to, 1050; K–12 assessment and, 1041; NCLB and, 1074; psychometrics and, 489; score reporting for, 917, 925; validation and, 445. *See also* state summative assessments

- summed score, 346, 696, 701
 Sun, A., 364
 Sun, M., 1147
 sunrise movement, 9, 30n5
 supervised LDA (sLDA), 548
 Suppes, P., 622–23, 741, 1522, 1577n5
 Surowiecki, J., 852
 survey administration, 1446, 1450, 1479, 1490
 survey congestion, 1480
 surveys, data collection with, 934–35
 Surveys of Enacted Curriculum (SEC), 1157, 1158, 1159
 Suter, L. E., 1431, 1488
 Svetina, D., 1365, 1456
 Svingby, G., 1031
 Swaak, T., 402
 Swaminathan, H., 61, 110n8, 411
 Swanson, D. B., 1302
 Swanson, L., 82, 585
 SWDs. *See* students with disabilities
 Swedish military interview, 1367
 Sweitzer, K., 1196
 Sweller, J., 842
 Swets, J. A., 839
 systematic variation, 165
 systemic evaluation, in validity argument, 229
 systemic racism, 11, 15, 384
 Syverson, S. T., 1195
- TACs. *See* technical advisory committees
 Takahashi, S., 1049
 Takane, Y., 664
 Takanishi, R., 139
 Takeuchi, K., 533
 Tal, E., 129, 656
 Talento-Miller, E., 1227, 1230, 1231
 TALIS. *See* Teaching and Learning International Survey
 Tamblyn, R., 1304
 Tannenbaum, R. J., 849, 854, 855, 867, 874, 908, 909
 Tao, W., 601
 Taras, H., 127
 target knowledge, accessibility to, 966
 target PLDs, 464, 465
 task design, for simulation-enabled performance tasks, 599
 task development, PAD and, 447
 task features, in domain analysis, 467
 task inventory approach, 1289–90
 task models, 245, 473, 1546
 Tate, R. L., 796
 Tatsuoka, K. K., 668
 tau-equivalent forms, 291, 292, 293, 294–95, 325, 349; true score and, 296–97
 Tauriz, G., 1342, 1343, 1381
 Taxonomy of Cognitive-Judgmental Tasks With Key Elements of Eight Standard-Setting Methods (and Two Variations), Including Paraphrased Statements of Cognitive-Judgmental Tasks, 835t
- Taylor, C. S., 121
 Taylor, J. J., 1072
 Taylor, K., 1281
 Taylor, R. D., 1346, 1350
 TBA. *See* technology-based assessment
 TCC. *See* test characteristic curve
 T-data measures, 1387
 teacher certification, 1281–83
 teacher cheating, 1148
 teacher evaluation, 1061, 1141–42, 1143, 1154
 teacher resource material, for professional development, 1036–37
 teacher working conditions, accountability testing and, 1147
 Teaching and Learning International Survey (TALIS), 1365, 1474
 Teaching Exercise and Fitness certification test, 871
 teaching expectations, sociocultural context relation to, 141
 Team Learning Process, 1040
 technical advisory committees (TACs), 217
 Technical Advisory Group, 1466
 technical manuals, reliability coefficients in, 362
Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA), 199, 200–201, 202, 204, 275n2
 Technology and Engineering Literacy assessment, NAEP, 475, 476, 594, 598
 technology-based assessment (TBA), 158, 255, 605–6, 608, 660, 1305; assessment design and, 583, 590, 635; background questionnaire and, 1490; comparability and, 610, 634–35; for consequential tests, 584–89, 600, 607, 631, 633–34, 636–37; construct fidelity and, 590, 592; construct-irrelevant variance in, 550, 601, 1320; decision-making and, 17, 582, 622, 630, 632–33; PBA compared to, 599, 600–603, 607; PISA and, 587, 592, 593, 594, 596, 603, 604; psychometrics and, 601, 604; response process and, 611–12, 635; validity and, 254, 260, 583, 634, 635
 technology-based reporting, 943–45
 technology-enhanced items (TEIs), 591f, 598, 601, 634, 635, 983–85; cognition and, 592; computer-based assessment and, 590–91; NAEP and, 587–88
 Technology-Enhanced Item Utility Framework, 984
 Teirstein, P. S., 1319, 1322
 TEIs. *See* technology-enhanced items
 Tellegen, A., 1378
 Teltemann, J., 117, 136
 Templin, J. L., 348, 350–51, 668, 1075
 Terman, L., 34, 36, 52, 54–55, 66, 85–86, 111nn12–13, 132, 385, 1523
 Terra Nova, 789, 986, 1034
 Terry, L., 360
 test accessibility. *See* accessibility
 test accommodations, 526, 794, 968, 969, 971–72, 1150; for accessibility, 609–10, 966, 972, 991; ADA and, 135, 139, 142, 971, 1283–86; construct-irrelevant variance and, 391, 516–17, 960; disabilities and, 135, 139, 141, 142–43, 167, 401–2; for ELs, 140, 143, 167, 517, 610, 974, 976; fairness and, 20, 145, 399, 424, 970, 997, 1363; flagging policies for, 405–6; for ILSAs, 1448–49, 1479; for inclusion, 143, 145; for licensure, 1280; for minority populations, 125; NCLB and, 142–43; sociocultural context for, 167–68; for special populations, 1230; state summative assessments and, 1048; for SWDs, 141, 145, 148, 165, 167, 254, 399, 401–2, 516, 969–70; systematic variation and, 165; TBA and, 254; test content and, 254, 991; test scores relation to, 214; validity and, 514–15, 516–18, 1165
 test administration, 17, 34, 514, 536, 963, 1288; CAT relation to, 1213; cheating and, 562; construct-irrelevant variance in, 398; COVID-19 pandemic and, 795, 803; on digital platforms, 154, 158; educational disparities relation to, 980; of ILSAs, 1424; linking relation to, 795, 800–801; for PISA, 1478; score reporting after, 925; test scores and, 515, 797; test security and, 518; UDA in, 973
 test administration modalities, validity and, 515–16
 test-based accountability, 21, 63, 226, 1146, 1155, 1169; flexibility waiver relation to, 1141–42; for SWDs, 1163–64, 1165–66
 test-based decisions, validity and, 221–22
 test-based inflation, 1160–61, 1162–63
 test-blind policy, 407, 423
 test characteristic curve (TCC), 283, 318, 745, 746, 747–48, 791
 test construction, 89, 199, 403, 1280, 1303, 1528
 test content, 13, 82, 236, 413–14, 737, 1291–92; fairness and, 387, 393–94, 399, 409, 963, 964, 989; PAD and, 910; panelists and, 828, 866, 868; preknowledge and, 519; standard setting and, 756, 849, 992; TBA and, 586; test accommodations and, 254, 991; test administration and, 514, 1288; test scores and, 137, 748; test security and, 1280, 1286; validity and, 200, 909; vertical scaling and, 790
 test development. *See* assessment development
 test equating, 58, 347, 766–67, 785; data collection and, 770, 775; subpopulation invariance evaluations and, 782–83
 test information function, 312, 316–17, 323, 329, 345, 348
 test interpretation, 970–71, 997, 999; construct validity and, 205, 206; IUA in, 208–12; score reporting relation to, 947;

- validation framework and, 205, 213; value judgments in, 206–7
- testlet model, 82, 598, 666–67, 1508n26
- Test of English as a Foreign Language (TOEFL), 233–34, 500, 540, 556, 585
- Test of English as a Foreign Language Internet-Based Test (TOEFL iBT), 534, 540–41, 556, 618
- test-optional policies, 407, 423; consequences of, 425; COVID-19 pandemic and, 21–22; for graduate school admissions, 1215; HSGPA relation to, 1196; minority populations and, 1195, 1196–97; test scores relation to, 1196, 1263n1; at UC, 1194; undergraduate admissions relation to, 1210
- test performance: construct validity and, 201; cut score relation to, 756; with interval properties, 755; norms and, 758; scale scores and, 738, 750; scaling and, 745; score reporting and, 915–16; SES relation to, 1471; test scores and, 749–50, 906; validity and, 1309; vertical scaling and, 791
- test preparation, 234, 396–97, 917, 967, 1050; accessibility and, 965; for admissions testing, 1243; preknowledge and, 1306; Preliminary Multistate Bar Review, 1286–87
- test purposes, score reporting relation to, 926
- test quality, funding relation to, 443–44
- test-retest reliability coefficient, 290–91, 364–65
- test scores, 50, 530; academic performance relation to, 426, 1224–25; for achievement tests, 200; for admissions testing, 1190, 1195; AI and, 532, 533; BARS and, 1375; civil life relation to, 129; coaching relation to, 1191; for collaborative problem solving assessments, 260; construct-irrelevant variance and, 207, 394, 399; construct validity and, 202–3; COVID-19 pandemic and, 4; for CR, 1454; for criterion-referenced tests, 338; critical interpretation of, 131; in CTT, 67, 1529; decision-making and, 215, 1508n27; equity relation to, 1167; extrapolation and, 249; fairness and, 19, 392; GPA and, 1263n3; GT and, 68, 213–14; IUA in, 197–98; KSAs and, 226–27; for large-scale testing, 194; MAP for, 674–75; medical school admissions relation to, 1214; multilevel construct validity and, 217–18; placement tests and, 1233; policy decisions relation to, 1116; public accountability with, 1138; Rasch model and, 1374; by raters, 521, 523; for reading comprehension, 1578n9; reliability of, 200, 279, 1297, 1301; response process and, 216; rubrics and, 550–51; on SAT, 1203–4, 1211; for selection tests, 233–34; sign-trait interpretations of, 449; for SJTs, 1384; test accommodations relation to, 214; test administration and, 515, 797; test content and, 137, 748; test-optional policies relation to, 1196, 1263n1; test performance and, 749–50, 906; validation framework and, 193–94, 199, 208, 210–12, 218, 1301–2; validity and, 389–90, 514, 1286, 1550; VAMs relation to, 1154, 1155. *See also* score reporting
- test scoring, 666, 685–86; construct-irrelevant variance and, 399; EAP in, 675; Lord–Wingersky algorithm and, 696; MAP and, 674; test administration relation to, 515
- test security, 17, 397, 1280; AIG and, 1305–6; CAT relation to, 1213; in continuous testing, 1307–8; for licensure, 1280, 1286–87; preknowledge and, 1306–7; for remotely administered exams, 402; TBA and, 589; validity and, 518, 1305; video cameras for, 516
- test specifications, 58, 207, 362, 475f, 747–48, 963; alignment of, 389; assessment design and, 456; assessment development and, 994; data collection and, 1289, 1290; domain analysis and, 455; linking and, 786; PAD and, 473–74, 475, 479, 492; panelists and, 858–59, 936; sensitivity guidelines in, 990–91; test content and, 992; TIMSS and, 975
- test specifications, RPLDs relation to, 474
- test-taker centered tasks, 828
- test-taker characteristics, 241
- test theory, 35, 38, 43; Spearman and, 34, 37, 42, 44, 47, 92; Thurstone and, 73. *See also* classical test theory; generalizability theory; item response theory
- test-to-standards alignment, 1159
- test use, 216; consequences and, 206, 221–22, 224, 228–29, 232, 233, 234, 235, 260; fairness and, 195–97, 218, 236, 400, 424–25; IUA and, 219; score interpretation and, 208–11, 217, 239; validity and, 192–94, 198–99, 205, 223, 245, 254, 261, 389–90. *See also* admissions testing
- Tetreault, J., 534, 535
- Tett, R. P., 1342
- text conversion, 526
- TF-IDF index, 547
- Thayer, D. T., 411–13, 773, 775
- theory-driven modeling, 616–17
- theory of action (TOA): for CBM, 1057; data use, 1034; evidentiary argument and, 486; of grade promotion exam, 1141; for interim assessment, 484, 484f, 485; IUA as, 219, 220–21; PAD and, 482
- Theory of Mental Tests* (Gulliksen), 65
- “theory of the construct,” 245
- The Politics of the National Assessment of Educational Progress (NAEP) Reading Framework Revision, 1565t
- thermometry, 1537–40
- theta-to-scale score transformation, 322
- think-aloud methods, test-taker characteristics relation to, 241
- Thissen, D., 81, 83, 112n26, 317, 318, 352, 362, 363, 660, 663, 667, 672, 675, 685–86, 694, 696, 704, 706, 719, 742, 747, 763, 794, 795, 802, 1360, 1421n4, 1522, 1532, 1533
- Thomas, C., 9
- Thomas, N., 680, 1452
- Thomas, R., 350
- Thomas, S., 1042
- Thomasson, G. L., 352
- Thompson, J., 1060, 1061, 1062
- Thompson, M., 1023
- Thompson, N., 1312
- Thompson, S. J., 244, 966, 967, 982
- Thompson, W. J., 348
- Thomson, W., 1538
- Thorndike, E. L., 36, 45, 53, 56, 58, 64, 73, 85, 90, 325, 363, 364, 386, 739–40, 756, 764, 1368, 1370, 1522, 1523–24, 1530, 1543, 1561
- Thorndike, R. L., 60, 421, 897
- Thorndike method, 325
- Thornton, A. E., 423
- 3PL model. *See* three-parameter logistic model
- three-parameter logistic model (3PL model), 667, 682, 707–8, 1561, 1578n10; ability estimates from, 747–48; interval properties and, 755; LOGIST and, 1534; Rasch model and, 1532, 1560; scaling and, 745; UIRT and, 704–5; vertical scaling and, 791
- threshold-signal detection judgments, 837, 839–40, 845
- through-year models, 632, 633
- Thudt, A., 912
- Thurlow, M. L., 19, 124, 135, 165, 235, 387, 398, 494, 511n2, 515, 516, 517, 597, 966, 967, 968, 969, 970, 971, 972, 981, 1018, 1056, 1164, 1166, 1168, 1230, 1280, 1283–84, 1357, 1449
- Thurstone, L. L., 36, 65, 71, 73–75, 199, 663, 739–40, 789, 1518–19, 1522, 1524, 1530, 1547
- Thurstone’s simple structure, 671
- Thurstonian IRT, 1379–80
- Thurstonian scaling, 791
- Tibshirani, R., 678
- Tiemann, G., 849–50
- Tienken, C., 1034
- Time* (magazine), 25
- time restrictions, 398, 1308
- Timperley, H., 626, 906, 943, 1022, 1023, 1025, 1026, 1031
- TIMSS. *See* Trends in International Mathematics and Science Study
- Tindal, G., 1055, 1056, 1059
- Tinker, M. A., 43

- Tippins, N., 1313
 Tisak, J., 350
 Tisak, M. S., 350
 TISM framework, 1524, 1535–36, 1562–63, 1564–65, 1566; reading comprehension and, 1516, 1541, 1542, 1543–44, 1545–49, 1555; scale scores and, 1558, 1560, 1561; scaling and, 1538–39; thermometry and, 1537–40; validity relation to, 1516–17, 1551–52, 1554
 Title I accountability requirements, 977, 1050, 1119
 Title VII, 1282
 TLI. *See* Tucker-Lewis index
 TOA. *See* theory of action
 Toch, T., 1121, 1141
 TOEFL. *See* Test of English as a Foreign Language
 TOEFL iBT. *See* Test of English as a Foreign Language Internet-Based Test
 Tomasello, M., 126
 “Tom W’s Specialty” experiment, 893n2
 Tong, Y., 116
 Top Five Estimated Proportions of Attribute Profiles in Diagnostic Classification Models, 711t
 topic models, 542, 547, 549–50, 555
 Topol, E. J., 1322
 Topping, K. J., 1030
 Torgerson, W. S., 71
 Torney-Purta, J., 1239
 Torres Irribarra, D., 71, 75, 93, 1577n2
 total survey error (TSE), 1445
 Toulmin, S. E., 450, 456, 482
 Toutkoushian, E., 1066, 1068–69, 1075
 Towles-Reeves, E., 970
 Town, R., 1375–76
 Toyama, Y., 156
 Trackman, 552
 training, 526; of ASR, 540; automated scoring and, 521, 536; for bias, 860, 861, 862t; for certification, 1278; for neural networks, 628; of panelists, 860–61, 1315; of raters, 522, 524–25
 trait-based assumptions, 450
 translation procedures: for ILSAs, 1442–43, 1472, 1479; for large-scale testing, 122; measurement error in, 1479
 transparency: in automated scoring, 622; human raters and, 554, 555; of LLMs, 619–20; in score reporting, 898; for validity, 550
 Traub, R. E., 286
Treatise on Electricity and Magnetism (Maxwell, J.), 1517–18
 treatment and control classrooms, 699–700
 Tremblay, K., 1238
 Trendler, G., 1553
 Trends in International Mathematics and Science Study (TIMSS), 90, 136, 138, 587, 589, 1340, 1507n3, 1507n10; accountability testing and, 1115; assessment design of, 1452; BARS and, 1458; BIB and, 1446; bilingual assessments and, 975; cognitive ability and, 1438; as computer-based assessment, 1427; data collection and, 1471, 1475; as digital assessments, 1455; educational policy relation to, 1468, 1490; education reform movements and, 139; IRT and, 1451; linking and, 794–95; population sampling and, 1443; PUF and, 1461–62; test accommodations for, 1449
 Trends in National Assessment of Educational Progress Reading Comprehension Grade 4 Students From 1998 to 2022, 1542f
 Trierweiler, T. J., 1232
 Trilling, B., 1348
 true scale score, variance of, 323
 true score, 67, 754; beta-binomial model and, 340; bivariate normal distribution and, 342; classification accuracy and, 338–39, 342; confidence intervals and, 750–51; congeneric forms and, 291–92; CSEM and, 325, 331; CTT and, 281–82, 286–87, 318, 1525, 1526, 1531; equating and, 765, 767, 778–79, 780; false negatives and, 1317; in IRT, 283; latent trait theory and, 1531; in Livingston and Lewis procedure, 343; raters relation to, 357; regression of, 764; scale-aligning methods and, 768–69; scaling and, 743; score weights and, 744; SEM and, 324, 366; tau-equivalent forms and, 296–97; universe score compared to, 1528; variance as, 345
 Trull, T. J., 1389
 Trumbull, E., 121, 140, 236, 237
 Trump, D. J., 25–27
 truncation, scale scores and, 751
 trustworthiness, 24, 1514–15
 TSE. *See* total survey error
 Tucker, B., 632, 765
 Tucker, C., 916, 935
 Tucker, L. R., 75, 77, 677, 1518
 Tucker, M., 1491
 Tucker, P. D., 1141
 Tucker equating, 765
 Tucker-Lewis index (TLI), 678, 697, 699, 708
 Tufte, E. R., 945
 Tukey, John W., 69, 660, 741, 753, 1532
 Tunstall, P., 1025
 Turek, M., 622
 Turing test, AES and, 550
 Turner, E. O., 228
 Turner, L. J., 1144
 Tversky, A., 838, 843, 845, 847
 2PL model. *See* two-parameter logistic model
 two-dimensional item factor model, 719, 1548
 two-level IRT, 712t, 713t, 715, 715t, 716t
 Two-Level Item Response Theory Growth Model With Condition Indicator as a Predictor: Data Structure, 715t
 Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Data Structure, 712t
 Two-Level Item Response Theory Growth Model With Treatment and Control Groups in Separate Data Columns: Factor Pattern, 713t
 two-parameter logistic model (2PL model), 665, 667, 704, 707–8, 721–22, 1548; ability estimates from, 747–48; conditional RP and, 682; fitted item factor model and, 719; graded response model relation to, 683; interval properties and, 755; Rasch model compared to, 1560; scaling and, 745; vertical scaling and, 791
 two-tier item factor model, 666–67
 Tyack, D. B., 149
 Tyler, J., 1035
 Tyler, Ralph, 89, 1118, 1205
 Tyler rationale, 1118
 Tyner, A., 1139
 Typical Standard Setting Workshop Process, 823f
 Tzamourani, P., 694, 696
 UAO. *See* universe of admissible observations
 UC. *See* University of California
 UC Consortium on Chicago School Research (UCCCSR), 1352–53
 UDA. *See* universal design
 UDL. *See* universal design for learning
 UG. *See* universe of generalization
 Uggerslev, K. L., 524
 UGPA. *See* undergraduate GPA
 Uijtdehaage, S., 1321
 UIRT. *See* unidimensional IRT
 Ulitzsch, E., 1480
Uncommon Measures (Feuer), 801–2
 undergraduate admissions, 1210–11, 1214. *See also* admissions testing
 undergraduate GPA (UGPA), 1224–25; academic performance relation to, 1232; GMAT relation to, 1227–28
 UNESCO Institute for Statistics, 1484
 unfolding approach, 1378–79
 unidimensional IRT (UIRT), 282, 320–21, 334–35, 345; 3PL model and, 704–5; adaptive testing and, 799–800; bifactor models *versus*, 708t; Bock–Aitkin EM and, 689; composite scores and, 346; CSEM in, 333; item analysis and, 703; MIRT and, 666; vertical scaling and, 792
 Unidimensional Item Response Theory Model Item Parameters, 705t
 Unidimensional Item Response Theory Model: Marginal X 2 and Standardized Local Dependence X 2 Statistics, 706t

- unidimensional pairwise preference task data, 1378
- Uniform Certified Public Accountant examination, 1294
- Uniform Guidelines on Employment Selection Procedures*, 1280, 1281, 1282
- unintended consequences, 220; of accountability, 1147–48, 1151; of test-based accountability, 1155
- United Nations' Sustainable Development Goals, 1435, 1484
- United States Medical Licensing Examination (USMLE), 586, 596, 599, 618, 653n2, 1229, 1285, 1289, 1291, 1294, 1338n2
- United States v. Fordice*, 149
- univariate generalizability theory (GT), 69, 298, 328–29
- univariate mixed model, 381n3
- universal design (UDA), 444; accessibility and, 973, 989; fairness in, 973; PAD and, 982, 994; TE and, 985
- universal design for learning (UDL), 244–45, 390–91; for accessibility, 394, 961, 973; in assessment development, 983, 1166; CLR and, 990; ELs and, 976; TE and, 985
- universe of admissible observations (UAO), 299–300, 301, 304–5, 381n3
- universe of generalization (UG), 300–304
- universe score, 197–98, 304, 1527; in GT, 67, 302, 1531; true score compared to, 1528
- University of California (UC), 133, 406–8, 1192–93, 1194, 1243
- University of California, Ethics, Compliance, and Audit Services, 1200
- University of Chicago, 73, 74, 81
- University of Copenhagen, 77
- University of Leipzig, 37, 46, 47
- University of Michigan, 1200
- University of North Carolina at Chapel Hill, 1200
- University of Pennsylvania, 10
- University of Texas, 1195, 1201, 1263n4
- Updated Evaluation Form for Reviewing Reports, 932t
- Ura, A., 1272
- Urban, C. J., 686
- Urban, W. J., 62
- Urban Data project, 1035
- user experience (UX), 498–99, 945
- USGAO. *See* Government Accountability Office, U.S.
- USMLE. *See* United States Medical Licensing Examination
- utilization inference, 275n4
- UX. *See* user experience
- Vacha-Haase, T., 355
- Valdés, G., 146, 1071
- Valencia, S. W., 463
- validation, 484; of admissions testing, 1199, 1201; assessment design and, 480, 723; of certification, 1322–23; CLR relation to, 966; consequences and, 232, 234; content-related evidence and, 393; design science and, 451; fairness and, 389; human raters for, 621; of ILSAs, 1437, 1449–50; interim assessment relation to, 250, 251–52; justice-oriented antiracist framework for, 996; learning progressions and, 1064, 1066–67, 1068; PAD and, 451, 481–82; of performance standards, 822; for PLDs, 489; RPLDs and, 463; rubric, 528; of summative assessments, 658; summative assessments and, 445; test-based accountability and, 1169; TISM framework relation to, 1516–17
- validation framework, 11–12, 17, 192; argument-based approach to, 215, 482–83; in assessment development, 199; cognitive activity in, 119, 131; consequences and, 221–22, 224; consequential evidence and, 197; ILSAs and, 1474; of reasoning, 129; test interpretation and, 205, 213; test scores and, 193–94, 199, 208, 210–12, 218, 1301–2; for workshop design, 856, 857t
- Validation Framework and Criteria for Performance-Level Descriptors, Design Patterns, and Item Development, 490t
- validity, 12, 13, 483, 1132, 1549; accountability and, 1188n1; of accountability policies, 1117; accountability testing and, 1114; of ACCUPLACER, 1237; achievement tests and, 201, 1193; of ACT, 1231; adaptive learning and, 630; admissions testing and, 21–22, 1223, 1224–25, 1225t; AES and, 255–57; AI and, 550; alignment and, 1156–57; argument-based approach to, 451, 909–10; assessment design and, 192, 194, 205, 208, 613; assessment for teaching and learning and, 1080; of assessment interpretation, 131, 192–93, 197, 204, 391; of automated scoring, 15, 255–56, 620–22; AVs and, 1372–73, 1374; BARS and, 1375–76; bias relation to, 132–33, 390, 1308; causal inferences and, 218–19; of civil service examinations, 384; coaching relation to, 396; cognitive-judgmental task and, 876; collaborative problem solving assessments and, 259; common formative assessments and, 1039; consequences relation to, 193, 222–24, 275n3; consequential, 1484, 1491; construct, 91, 198, 200–204, 205, 206, 208, 217–18, 1474; construct-irrelevant variance and, 227, 282, 597–98; construct-relevant variance and, 610–11; content, 200–201, 202, 203, 585, 1237; correlation coefficient and, 233; COVID-19 pandemic and, 795; in credentialing, 23, 1269, 1273, 1303, 1309, 1323; criterion, 199–201; criterion-related, 449, 556; cultural, 992–93; curricular, 150–51; of data, 527; data story and, 911–12, 914; decision/use inferences and, 215; differential, 419, 423; in digital environments, 137; for diverse groups, 192; ECD and, 551, 1292; equity and, 235; expected error variance relation to, 283; fairness and, 14, 34–35, 168, 193, 194–97, 229, 275n1, 387–89, 393, 408–9, 424, 552–53, 608, 724, 960, 1351; forced-choice methods and, 1380–81; GPA and, 214; of grading, 1041, 1042–43; of GRE, 1228; GT and, 197–98; of ILSAs, 1476, 1481; innovative item types and, 1294; of interim assessment, 252, 1033–34; IUA and, 193, 203–4, 206, 208, 212, 219, 228, 238, 251, 253, 260, 928; learning progressions and, 1064, 1065; LLMs for, 637; of LSAT, 1232; of machine learning, 538; of machine scoring, 554, 555; of MCAT, 1230, 1232–33; Messick's Facets of, 205f; of milestone ratings, 1321; nomological, 203; PAD and, 470, 494; in performance assessment, 220; performance inferences and, 212; PISA and, 1437; of placement tests, 1234, 1236; preknowledge relation to, 1306; raters and, 521–22, 1368; reliability and, 194, 195, 197; of remotely administered exams, 400–401, 402; RSM and, 1362; of score inferences, 245, 514; of score interpretation, 125, 192–94, 197, 207, 210, 212, 217, 218, 223, 224, 239–40, 960, 961, 1280, 1292, 1305, 1308, 1553; score reporting and, 896, 897, 903, 906, 907–9, 937, 947; score weights and, 743; situated learning theory and, 244; SJTs and, 1382, 1385–86; sociocultural context and, 119, 125; sociocultural learning theories and, 261; standardization relation to, 124–25, 258, 514–15; standard-setting design and, 877; of summed score, 701; TBA and, 254, 260, 583, 634, 635; of teacher certification, 1283; test accommodations and, 514–15, 516–18, 1165; test administration and, 15, 514; test-based decisions and, 221–22; of test interpretation, 209; test scores and, 389–90, 514, 1286, 1550; test security and, 518, 1305; test use and, 192–94, 198–99, 205, 223, 245, 254, 261, 389–90; TISM framework relation to, 1516–17, 1551–52, 1554; of Trackman, 552; training relation to, 861. *See also* predictive validity
- validity argument: admissions testing and, 231; consequences and, 229–30, 389; for culturally responsive assessment, 235, 237, 246, 247t, 248t; IUA and, 234; PAD and, 473; systemic evaluation in, 229
- validity evidence, growth model analyses and, 230

- Validity of Undergraduate and Graduate Admissions Tests for Predicting First-Year Grade Point Average (GPA), 1225t
- validity research, for selection procedures, 59
- validity responses, raters and, 524
- validity theory, 12, 66, 198, 199; consequences and, 232; in K–12 assessment, 194, 229, 249–50; validation framework compared to, 193
- Valtcheva, A., 1031
- value-added models (VAMs), 1390; accountability in, 140; bias and, 351; confidence intervals and, 366; high-stakes testing and, 1155; reliability in, 350; test scores relation to, 1154, 1155
- value judgments, 206–7, 837, 840, 1318–19
- VAMs. *See* value-added models
- van Aken, J. E., 451
- Van Ameringen, M., 127
- Vanchu-Orosco, M., 974
- van den Bosch, R. M., 1055, 1057, 1059
- van den Heuvel, J. R., 972
- van der Ark, L. A., 1532, 1551–52
- Vandergrift, J. L., 1304, 1309
- Van der Kleij, F. M., 908, 1025
- van der Linden, W., 77, 80, 354, 519, 615, 683, 749, 765, 768, 797, 798, 1306, 1360, 1361, 1380, 1459, 1529–30, 1548
- Vandermeulen, N., 631
- van der Ploeg, A., 1037
- van der Vleuten, C. P. M., 1321
- van de Vijver, F. J. R., 119, 147, 155, 1365, 1436, 1456–57, 1479
- van Dijk, T. A., 1544
- Van Dunk, E., 1134
- van Ewijk, R., 402
- VanLehn, K., 623, 627
- Vanlommel, K., 1045
- Van Meer, E., 622–23
- Van Moere, A., 523
- van Rijn, P. W., 1063
- Vantage Learning, 559
- van Zundert, M., 1030
- variance, 66–67, 68, 326, 329; AI and, 358; application of analysis of, 297–98, 305; bias and, 312–13; in CTT, 302–3, 325; empirical ability distribution and, 353; in G studies, 299, 300–301, 302, 307, 309; in GT, 360; latent proficiency, 311, 314; in model-based observed scores, 319; in rater reliability, 356–58; raw scores and, 320; reliability and, 69; score weights and, 744; of true scale score, 323; as true score, 345. *See also* error variance
- variance algebra, 714
- Variance and Covariance Components for the Extended Multivariate Design Example, 308t, 309t
- variance-covariance matrices, 773–74
- variance-ratio definition, 317; error variance and, 288–89; MLE compared to, 316; reliability coefficients and, 289, 315
- Variance-Ratio Reliability Coefficients for the Maximum Likelihood Estimator, 315t
- Various Characterizations of Cognition Related to Assessment, 457t
- Varsity Blues, 22, 30n8, 397, 1198
- Vaswani, A., 549
- Vaughn, K. W., 113n32
- Vazire, S., 1368
- Vazquez, E., 1474
- Vedel, A., 1362
- Veldkamp, B. P., 583
- Venator, J., 1345, 1351
- Ventura, M., 493
- Vera, H., 129
- Vera-Toscano, E., 1471
- Verbert, K., 924
- Verheggen, M. M., 870
- Verhelst, N., 356–57
- Verhoeven, B. H., 870
- Verschelden, C., 127
- Verstralen, H., 356–57
- vertical scaling, 755, 789–90; adaptive testing relation to, 799; IRT and, 791–92
- Vestman, R., 1342, 1343, 1366, 1367, 1369
- Vetro, A., 622
- Vevea, J. K., 1296
- The Victorian Quality Council, 1138
- video cameras, 516, 552
- video prompts, 528, 529, 530f
- Viegut, D., 1038, 1039
- Vigdor, J., 1144
- Vinokur, A., 831, 863
- Vinovskis, M. A., 1131–32
- virtual agents, 259
- Virtuous Spiral, 1435f, 1438, 1442
- Virvou, M., 626, 629
- Visible Learning initiative, 903
- Visual Representation of a Sociocultural Theory of Reading Comprehension From the Public Comment Draft of the Reading Framework for the 2026 National Assessment of Educational Progress, 1545f
- Vo, T. T., 8–9
- vocabulary, in assessment design, 395
- Volante, L., 1468, 1484
- Voluntary Framework of Accountability, 1240
- voluntary professional credentialing, 1276–77
- Voluntary System of Accountability, 1240
- von Collani, G., 1388
- von Davier, A. A., 946, 983, 1072, 1075, 1433, 1436, 1437, 1442, 1450, 1451, 1452, 1453–54, 1455, 1458, 1459, 1477, 1479, 1482–83, 1508n32, 1566
- von Davier, M., 75–76, 82, 137, 255, 258, 603, 604, 660, 668, 669, 679, 765–66, 770, 774–75, 777–78, 781, 782, 802, 1296, 1320, 1365, 1378
- Vukasovic, T., 1346
- Vygotsky, L. S., 116, 126, 992, 1544
- Wackerle-Hollman, A., 976
- Wagemaker, H., 1467, 1469
- Wagenaar, R., 1239
- Wainer, H., 12, 57, 81–82, 92, 112n26, 347, 362, 363, 585, 598, 666–67, 675, 676, 685–86, 696, 707, 719, 901–2, 903, 931, 934, 938, 945, 1296, 1308, 1421n4
- Wainer, M. E., 404, 408
- waiver accountability policies, 1125–27
- Waldren, S. E., 1320
- Walker, A. A., 520
- Walker, G. F., 110n1
- Walker, M. E., 610, 766, 767, 784, 795
- Walker, W. E., 116
- Wall, D., 234
- Wallace, A. R., 37–38
- Wallace, T., 1059
- Wallach, H. M., 549
- Wallin, G., 801
- Walsh, P., 366
- Walter, M., 1118
- Waltman, K. K., 1042
- Wan, L., 341, 592, 1298
- Wand, J., 1371, 1372
- Wang, A., 549, 1232
- Wang, C., 353, 366, 539
- Wang, J., 525, 789
- Wang, S., 515
- Wang, T., 345, 602, 775, 779, 798
- Wang, W.-C., 1379
- Wang, X., 541
- Wang, Z., 357
- Wantman, M. J., 613
- Wanzer, D., 1353
- Warburton, 1387, 1388
- Ward, W. C., 585
- Warm, T. A., 315
- War on Poverty, 977
- Warren, B., 152
- Warren, J. R., 1148, 1149
- Warrington, 353–54
- Watanabe, T., 407
- Waters, B. K., 898
- Watson, J. B., 1543
- Watts, T. W., 1388
- Way, W. D., 584, 585, 586–87, 601, 605, 608, 609, 855, 1305, 1310, 1311
- Wayne, D. B., 870, 874
- WCAG. *See* Web Content Accessibility Guidelines
- wealth, 979–80, 1209, 1210
- WEALTH scale, 1478
- Webb, M. L., 164
- Webb, N. L., 459, 493, 681–82, 1157, 1159
- Webb, N. M., 298, 1528
- Webb approach, 1157, 1158, 1159
- Webber, C., 384, 385

- Web Content Accessibility Guidelines (WCAG), 444, 495
- Weber, E., 46–47, 71, 1518
- Weber-Fechner law, 46–47
- Webster, L., 81
- Weeks, J. P., 789, 791, 792, 793
- weighted deviation method, 82
- Weinberger, C. J., 1344
- Weiner, J. A., 516
- Weinland, T. P., 53, 54
- WEIRD. *See* Western, educated, industrialized, rich, and democratic
- Weirich, S., 1388
- Weisberg, D., 1141
- Weiss, D. J., 60, 584, 585
- Welch, B. L., 41
- Welch, C., 119
- Welch, C. J., 470, 472, 473
- Wells, A. S., 1135
- Welsh, M. E., 1043
- Wendler, C., 1213, 1219, 1230
- Wendt, A., 1293
- Weng, Q., 1304, 1384
- Wenger, E., 127, 1544
- WER. *See* word error rate
- Werts, C. E., 417
- Wertsch, J. V., 116, 126
- West, M. R., 1150, 1340, 1354–55, 1362–63, 1366, 1370
- Western, educated, industrialized, rich, and democratic (WEIRD), 161–62
- Westring, P. A., 1043–44, 1209, 1225, 1226
- Westring, A. J. F., 1384
- Wetzel, E., 1452
- What Works Clearinghouse, 1345
- “what you see is all there is” (WYSIATI), 838, 846
- Wheeler, Jordan, 17, 542, 547
- Whelpley, C. E., 1384
- Whetzel, D. L., 1382
- Whipple, G. M., 386
- White, H., 677, 1021
- White, P. A., 1033
- White, W. D., 1266
- Whitney, D. R., 1190
- Whittington, D., 355
- Wiberg, M., 765–66, 773, 801
- Wickham, H., 945
- Wiener, D., 970
- Wigfield, A., 127
- Wiggins, G., 1028
- Wightman, L. E., 765, 772
- Wilbur, W. J., 547
- Wilcox, K. T., 548
- William, D., 250, 1020–21, 1023, 1031, 1034, 1041, 1063
- Wilkerson, S. B., 1036, 1041, 1050
- Wilkinson, L., 945
- Willett, J. B., 350, 1152
- Williams, 1242
- Williamson, D. M., 255, 256, 550, 551, 618, 620, 653n4, 1319
- Williamson, J. A., 995
- Williamson, V. S. L., 802
- Willingham, W. W., 425, 964, 968, 1043, 1191, 1199
- Wills, A. J., 839
- Wilner, J., 407
- Wilson, M., 148, 156, 237, 252, 357, 445, 611, 662, 681, 723, 1064–65, 1066, 1067, 1068, 1069, 1070, 1072, 1073, 1075, 1076, 1533, 1535, 1551–52
- Wilson, Woodrow, 52
- Wind, S. A., 524, 1533
- Wine, J. D., 127
- Wingersky, M. S., 318, 321, 696, 765, 778, 796, 1534
- Winne, P. H., 1025
- Winstead, L., 134
- Winstone, N. E., 1025, 1027
- Winter, P. C., 120, 241, 602, 981
- Winters, M. A., 1141, 1144, 1149
- Wirth, R. J., 665
- Wirtz, W., 1433
- Wise, L. L., 390, 632, 749
- Wise, S. L., 444, 498, 1240, 1388, 1480
- Wiseman, A., 1486–87
- Wisniewski, B., 943
- Wissler, Clark, 42–43, 51
- Wolf, M. G., 678, 701
- Wolf, M. K., 145, 146
- Wolf, T. A., 48
- Wolfe, E. W., 521, 522, 524, 525, 536, 551, 983
- Wolf-Yadlin, R., 423
- Wollack, J. A., 515, 518, 519, 520, 897, 1307
- Wong, J., 1076
- Wong, M., 1143–44
- Wood, B. D., 45
- Wood, D. A., 73
- Wood, S. W., 366
- Woodruff, D. J., 325, 338, 361
- Woods, C. M., 663
- Woods, S. A., 1371
- Woodworth, R. S., 1524
- Woolf, B., 626
- Woot Math, 623
- word2vec, 549
- word error rate (WER), 539, 579n2
- workforce, 133, 136
- working model, 677
- workplace outcomes, 1342, 1369
- workshop design, validation framework for, 856, 857t
- workshop process submodel, 829, 832, 856
- Works Progress Administration, 75
- World Book Company, 86
- Wrabel, S. L., 1123, 1124, 1125, 1126
- Wright, B., 78, 79–80, 81, 112n25, 740, 1368, 1370, 1533, 1534
- Wright, L., 241
- Wright, N. K., 742, 747, 801
- Wright map, 927, 1065
- writing prompts, justice and, 995–96
- written testing, 84, 121–22
- Wry, E., 1436–37
- Wu, C., 361, 549
- Wu, C. J., 689
- Wu, H., 1365, 1374
- Wundt, W., 43, 47, 51
- Wu Y.-C., 969, 980–81
- Wylie, E. D., 1028, 1029
- Wynne, M., 1239
- Wyse, A. E., 225, 346, 347, 868, 870, 874
- WYSIATI. *See* “what you see is all there is”
- Xi, N., 794
- Xi, X., 541
- Xiao, Y., 927–28
- Xie, S., 541, 556
- Xin, T., 313, 319, 360
- Xiong, H., 547
- Xiong, J., 542
- Xiong, W., 526
- Yamamoto, K., 1447–48, 1451, 1453, 1455, 1459
- Yan, Z., 1029
- Yaneva, V., 137
- Yanez, E., 992
- Yang, J. S., 720
- Yang, Y., 349
- Yao, L., 357
- Yao, S.-Y., 1069
- Ye, L., 1452
- Yeager, D. S., 425, 1026, 1030, 1137, 1345, 1348
- Yeld, N., 873
- Yen, Wendy, 81, 89, 278, 656, 747
- Yen, W. M., 791, 920
- Yerkes, Robert, 52–53, 54, 56, 764, 1522
- Yeung, C. K., 628
- Yeung, D. Y., 628
- Yeung, S. F., 520
- Yin, P., 783, 788
- Yoakum, C. S., 53, 54, 764
- Yoon, G. E., 26
- Yoon, S. Y., 541
- Young, J. W., 120, 517, 1219, 1231
- Young, R. E., 553
- Young, V. M., 408, 419, 423, 493
- Yoyou, W., 1390
- Yu, A., 122
- Yu, L., 605
- Yule, G. U., 43, 1526
- Zaglaniczny, K. L., 1290
- Zaki, M. J., 549
- Zamarro, G., 1388
- Zamudio-Suarez, F., 1236
- Zantoni, W., 1390

- Zapata-Rivera, D., 623, 902, 910, 924, 925,
928, 930, 931, 937, 942–43, 944, 1296
- Zapf, A., 357, 360
- Zechner, K., 538, 540, 541, 556
- Zedeck, S., 1214, 1355, 1374, 1375
- Zenderland, L., 56
- Zeng, W., 1018
- Zenisky, A. L., 19, 123, 517, 591, 736, 799,
896–98, 901, 902, 905, 906, 907, 908,
912, 914–15, 916, 918–19, 923, 927, 928,
929–32, 934–37, 943, 944, 945, 946–47,
961, 984, 1049, 1295
- Zesch, T., 749
- Zhai, S., 549
- Zhai, X., 258
- Zhang, H., 35
- Zhang, M., 256, 595, 614, 617, 618, 620–21,
631, 637
- Zhang, W., 781, 782
- Zhang, Y., 164, 352, 353
- Zhang, Z. M., 549
- Zheng, G., 1073
- Zhou, E., 1213
- Zhou, Y., 1303, 1483
- Zickar, M. J., 1360
- Zieky, M. J., 119, 414, 424, 861, 910, 960, 966,
967, 983
- Zimmerman, B. J., 1022
- Zimmerman, J., 1229
- Zinnes, J. L., 741
- Zlatkin-Troitschanskais, O., 1238
- Zu, J., 23, 387, 962, 1115, 1384
- Zu Eissen, S. M., 520, 521
- Zumbo, B. D., 217, 218, 254, 350, 1481
- Zwick, R., 15, 207, 225–26, 235, 240,
387, 392, 393, 396, 398, 404, 405, 410,
412, 413, 418, 423, 425–26, 438n9,
494, 515, 516, 632, 931, 960, 961, 962,
964, 999, 1018, 1164, 1168, 1188n1,
1198, 1200, 1202, 1209, 1214,
1230, 1242, 1280, 1281, 1357, 1449,
1547, 1558

